Biomedical Significance of Collagen Glycosylation – With Focus on the Collagen Glucosyltransferase

---

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Jürg Andri Cabalzar

von

Ilanz GR

Promotionskomitee

Prof. Dr. Thierry Hennet (Vorsitz)

Prof. Dr. Matthias Baumgartner

PD Dr. Lubor Borsig

Prof. Dr. Arnold von Eckardstein

Zürich, 2015

TABLE OF CONTENTS

SUMMARY

Collagen is ubiquitous in the human body being the main component of connective tissues such as cartilage, bone, tendon, and ligaments. Disorders of connective tissues often originate from genetic defects in collagen genes or genes encoding collagen modifying enzymes. Collagen encompasses a superfamily of glycoproteins carrying extensive post-translational modifications. Defective post-translational modifications of collagen lead to severe developmental disorders with mild to lethal outcomes emphasizing the importance of these modifications. The oxidative modification of proline and lysine residues is well described and important for the functionality of collagens. Oxidatively modified lysine residues are further modified with glycan molecules, a process called collagen glycosylation. Eight decades ago, the existence of Glucose-Galactose-Hydroxylysine collagen glycans were discovered and subsequently found to be conserved in structure from sponge to human. The GLT25D1 and D2 galactosyltransferases that initiate these glycans were first characterized and reported only recently. The second enzyme catalyzing the elongating reaction, i.e. the transfer of glucose to galactosylhydroxylysine has yet to be identified, which is an aim of the present work. Application of a conventional purification approach followed by protein identification by tandem mass spectrometry repeatedly identified the putative glycosyltransferase UGGT2. Recombinantly expressed UGGT2 did not, however, possess a collagen glucosyltransferase activity. Biochemical approaches to identify the glucosylating enzyme did not provide clear answers whereas a bioinformatics approach revealed the enzyme GTDC1 that should be considered a strong candidate for the collagen glucosyltransferase.

Association of heritable developmental disorders with defects in collagen modifying enzymes suggests that defects in the glycosylating enzymes may be found in patients with connective tissue disorders. Screening human patient cells from several untyped connective tissue disorders did not reveal defects in collagen glycosylating activities. Abnormalities of collagen in many human diseases make collagen a central molecule in medical research. Apart from medical interest, there is a large demand for, and limited supply of recombinant collagen for biotechnological applications in tissue engineering, tissue remodeling after transplantation, and wound healing. The inability of the human hydroxylating and glycosylating enzymes to efficiently modify large amounts of collagen has hampered the biotechnological production of collagen. With the identification of two collagen modifying enzymes from *Acanthamoeba polyphaga* mimivirus which possess enzymatic activity in bacteria, a system for the production of high yield recombinant collagen was established. The coexpression of the two mimiviral hydroxylases with human collagen in bacteria enabled efficient proline and lysine hydroxylation of collagen close to physiological levels without the need to supplement with cofactors. The high yield from bacterial expression combined with a high degree of prolyl and lysyl hydroxylation provides the framework

for the large-scale production of post-translationally modified recombinant collagens for human applications.

Conclusively, the work conducted in this thesis provides new insights in the biology of collagen glycosylation and describes a new system for the production of physiological collagen for biotechnological applications.

ZUSAMMENFASSUNG

Kollagen ist der Hauptbestandteil der extrazellulären Matrix und allgegenwärtig im menschlichen Körper. Bindegewebe wie Knorpel, Knochen, Sehnen und Bänder erhalten ihre Form und Festigkeit hauptsächlich durch ihren Kollagengehalt. Eine der häufigsten Ursachen von vererbbaren Bindegewebskrankheiten sind genetische Defekte, welche zu Kollagenmangel oder zu fehlerhaften Kollagenstrukturen und schlussendlich zu schwerwiegenden Entwicklungsstörungen führen. Oftmals wurden Defekte nicht nur in den Genen von Kollagen gefunden, sondern auch in Genen Kollagen modifizierender Enzyme. Der Krankheitsschweregrad, ausgelöst durch die fehlerhaften Modifikationen, variiert von leicht bis tödlich und verdeutlicht die Wichtigkeit der Kollagenmodifikationen. Die oxidativen Modifikationen von Prolin- und Lysinresten sind sehr wichtig für die Funktionalität von Kollagen und bereits beschrieben. Oxidativ modifiziertes Lysin wird mit Galaktose und Glukose glykosiliert und bildet somit die Kollagen spezifische Glykanstruktur. Obwohl die Kollagenglykosilierung schon vor acht Jahrzehnten erstmals beschrieben wurde und die Struktur von Hydroxylysin gebundener Galaktose-Glukose-Dimeren im gesamten Tierreich konserviert ist, wurde erst vor sechs Jahren das erste der für die Glykosilierung verantwortlichen Enzyme entdeckt, die Kollagengalaktosyltransferase. In der vorliegenden Arbeit wurde die Identifikation des zweiten Enzyms, der Kollagenglukosyltransferase, mittels herkömmlicher Proteinaufreinigung und anschliessender Proteinidentifizierung durch Massenspektrometrie angestrebt. Die vermeintliche Glykosyltransferase UGGT2 wurde wiederholt identifiziert, konnte aber nach spezifischen Aktivitätstests nicht als Kollagenglukosyltransferase bestätigt werden. Da die biochemischen Ansätze keine klaren Resultate bezüglich der Identifizierung der Kollagenglukosyltransferase ergaben, wurde versucht das Enzym mittels bioinformatischen Methoden zu identifizieren. Daraus resultierte das Enzym GTDC1 als möglicher Kandidat, welches unbedingt auf seine Aktivität als Kollagenglukosyltransferase getestet werden sollte.

Die häufig diagnostizierten Kollagenfehlbildungen und -mangelerscheinungen in Bindegewebskrankheiten werden oftmals mit Defekten in den Kollagen modifizierenden Enzymen assoziiert. Bis dato wurde kein Fall beschrieben, bei welchem spezifisch die Glykosyltransferasen betroffen sind. Sämtliche von uns analysierten Zellen aus Patienten mit unbekannten Bindegewebskrankheiten verfügten über vollständig funktionale Glykosyltransferasen. Kollagenabnormalitäten in unterschiedlichen humanen Krankheiten machen Kollagen zu einem zentralen Molekül der medizinischen Forschung. Nebst dem medizinischen Interesse an Kollagen, besteht auch eine Nachfrage an rekombinant hergestelltem Kollagen für biotechnologische Anwendungen im Rahmen der Gewebeherstellung und -modellierung nach Transplantationen oder in der Wundheilung. Die Herstellung von

physiologischem Kollagen war bis anhin erschwert, weil die menschlichen Hydroxylasen in auf grosse Mengen ausgerichteten Expressionssystemen nicht aktiv sind. Mit der Entdeckung von zwei *Acanthamoeba polyphaga* Mimivirus Kollagenhydroxylasen, welche in Bakterien aktiv sind, haben wir ein Expressionssystem für die Herstellung von grossen Mengen an rekombinantem hydroxyliertem Kollagen geschaffen. Die Coexpression dieser zwei viralen Hydroxylasen mit humanem Kollagen in Bakterien ergab eine effiziente Hydroxylierung von Prolin- und Lysinresten annährend den physiologischen Werten. Die grossen Mengen Kollagen aus bakterieller Expression und der hohe Grad an Hydroxylierung schaffen die Rahmenbedingungen für eine gross angelegte Produktion von rekombinantem Kollagen für biotechnologische Anwendungen.

Die Resultate der vorliegenden Dissertation liefern neue Erkenntnisse für die Beschreibung der Kollagenglykosilierung und dokumentieren ein neues System für die Produktion von physiologischem Kollagen für biotechnologische Anwendungen.

LIST OF ABBREVIATIONS

ADAMTS	A disintegrin and metalloproteinase with thrombospondin motifs 2
BMP-1	Bone morphogenetic protein 1
CAZy	Carbohydrate active enzymes
COL1A1	Alpha-1-chain of collagen type I
ColGalT	Collagen galactosyltransferase
ColGlcT	Collagen glucosyltransferase
ConA	Concanavalin A
CRTAP	Cartilage-associated protein
Cys	Cysteine
DEAE	Diethylaminoethanol
DXD	Aspartic acid – any amino acid – aspartic acid
ECM	Extracellular matrix
EDS	Ehlers-Danlos syndrome
ER	Endoplasmatic reticulum
FCS	Fetal calf serum
Gal	Galactose
Glc	Glucose
Gly	Glycine
GT	Glycosyltransferase
GTDC1	Glycosyltransferase-like domain-containing protein 1
Hyl	Hydroxylysine
Hyp	Hydroxyproline
LH	Lysyl hydroxylase
LOX	Lysyl oxidase
NaCl	Sodium chloride
OI	Osteogenesis imperfecta
PBS	Phosphate buffered saline
PLOD	Procollagen-lysine, 2-oxoglutarate 5-dioxygenase
Pro	Proline
P3H	Prolyl 3-hydroxylase
P4H	Prolyl 4-hydroxylase
PPIB	Peptidyl prolyl cis-trans isomerase B
PTM	Post-translational modification
Ser	Serine
TBS	Tris buffered saline
UGGT	UDP-glucose:glycoprotein glucosyltransferase
UDP	Uridine-diphosphate

INTRODUCTION

Collagen is the most abundant structural protein in vertebrates. One of the critical factors for the structural and biochemical functions of collagen is post-translational modification (PTM). Since molecular research on collagen biosynthesis started in the 1930s, PTMs and their contribution to collagen architecture were extensively studied. Hydroxylation of proline plays a significant role in collagen folding and stability by enhancing the collagen triple helix thermal stability [1]. Lysine hydroxylation is evidentially important for proper crosslink formation between collagen fibers [2, 3]. However, little is known about collagen glycosylation. Even though the structure of collagen glycosylation is conserved in the animal kingdom from sponge [4] to human [5], the function of collagen glycosylation remains elusive. The genes coding for the collagen core glycosyltransferases have been discovered only in the past years [6]. The genes coding for the elongating glycosyltransferases have not been identified to date.

Defects in several collagen modifying enzymes lead to developmental disorders and fragility of connective tissue and bones with various severity. Until now, defects in conjunction with collagen glycosylation have not been described. Pathologic high levels of modification derived from over hydroxylation and over glycosylation result in mineralization defects in bones [7, 8]. However, the cause of the over modification does not come from defective glycosylating enzymes but rather from kinetic folding differences of defective collagen chains leading to prolonged enzymatic activity. High levels of PTMs do not infer a pathological sign per se because the degree of modification is highly diverse among tissues and collagen types. The molecular mechanism regulating type, amount, and position of the PTM is not known. Until now, 23 genes encoding various collagen modifying enzymes exist in human [9]. However, the functional contribution of the modifications among different cell types and tissues is not fully described for all modifications. In particular the function and various extent of collagen glycosylation remains to be studied.

COLLAGEN STRUCTURE AND BIOSYNTHESIS

Among the genomes of vertebrates (and higher invertebrates) are 28 distinct collagen types encoded by at least 45 genes. The collagen types are classified predominantly according to domain structure and their suprastructural organization [10]. Some collagens form fibrils with a characteristic periodicity of the forming units giving a striated pattern. They are grouped as fibril-forming collagens. Other collagens associate with collagen fibrils and are grouped as fibril-associated collagens with interrupted triple helices (FACIT). Further groups are network-forming collagens, transmembrane collagens and molecules with collagenous domains. For simplicity, this chapter covers only the biosynthesis of the most abundant collagen fibrillar type I collagen.

Composition of fibrillar type collagen

Collagen is a large trimeric protein. Its subunits comprise three α-helical strands coiling together into a right handed triple helical suprastructure. The three α-helical strands can either be identical or different forming homotrimeric or heterotrimeric structures, respectively. The prerequisite for its characteristic structure is encoded in the collagen sequence. Every collagen amino acid sequence contains a collagenous motif with a glycine (Gly) residue at every third position. Repeats of the Gly-Xaa-Yaa motif form the characteristic domains of collagen molecules. The high abundance of Gly is necessary since only the small hydrogen side group of Gly fits in the center pocket of the triple helix. Every other amino acid than Gly has a bulkier side group, hence, any mutation occurring at the Gly position will likely have dramatic consequences. Gly substitutions have been identified in Osteogenesis imperfecta (OI) patients and Ehlers-Danlos syndrome (EDS) patients resulting in structural abnormality of the collagen helix or defective triple helix formation (see below) [11, 12].

The side chains of the remaining Xaa and Yaa amino acid positions protrude towards the outside of the triple helix. This arrangement allows the accommodation of any amino acid but often proline (Pro) or hydroxyproline (Hyp) are found. The amino acid composition of the α1-chain of human collagen type I contains 26.7% Gly residues and 19% Pro residues of which about 45% are hydroxylated [13]. The presence of large amounts of Pro and Hyp together with the recurrent Gly give rise to an alpha helical conformation. The steric repulsion of the pyrrolidine rings from Pro residues results in larger axial distance between each amino acid compared to normal alpha-helices. Consequently, the helix structure is a stretched polyproline helix with restricted mobility due to the pyrrolidine rings. These rigid pyrrolidine rings have stabilizing effects since they limit rotation around the peptide N-C bond. Three polyproline α-helices provide the subunits for the triple helical procollagen chain. The triple helical collagenous domain of every procollagen is flanked by non-collagenous N-and C-terminal domains. In fibrillar collagen types these non-collagenous domains are called propeptides and are cleaved off prior to fibril formation (Figure 1). In contrast, in network forming collagen types and FACIT collagen types the propeptides remain uncleaved and significantly contribute to the assembly of the collagen supramolecular structure [9].

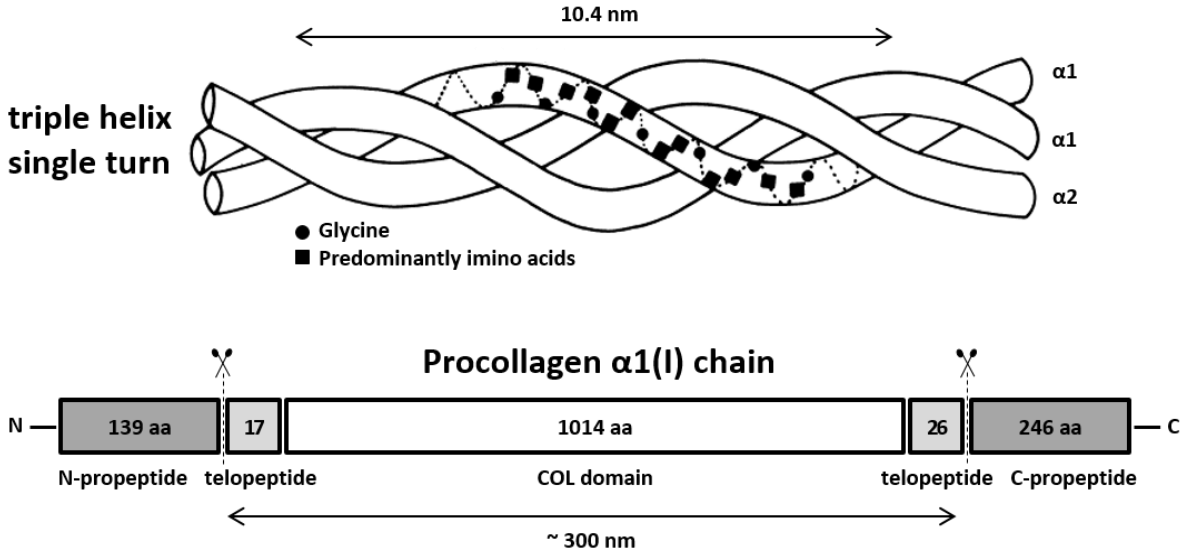


Figure 1| **Schematic representation of fibrillar type I collagen.** The α1-chain of procollagen type I consists of a 1'014 amino acid (aa) long collagenous domain flanked by telopeptidyl regions, a N-propeptide, and a globular C-propeptide. Upon collagen maturation the propeptides get cleaved off and the triple helical collagenous domain (COL domain) remains. Two procollagen α1(I) and one α2(I) chain assemble to a triple helical structure with glycine facing the center pocket. Figure adapted from *(Nimni ME, Harkness RD: Molecular structure and functions of collagen. In Nimni ME (ed): Collagen. Vol 1. Boca Raton, CRC Press, 1988).*

Triple helix assembly and fibrillogenesis

Collagen molecules are assembled from three monomeric pro-α-chains. Each α-chain gets synthesized into the rough endoplasmatic reticulum (ER) where it gets hydroxylated and glycosylated prior to folding and assembly into the procollagen triple helix. The extent of modification depends on the helix forming kinetic, which is in turn affected by the polypeptide primary structure. In general, fibrillar collagen types carry fewer modification compared to network forming types. The chain selection requires high specificity prior to trimerization and triple helix formation in order to avoid unfavorable α-chain trimers or multimeric aggregates with short misaligned triple helical patches. For example, pro-α1(I) chains can form homotrimers or heterotrimers with pro-α2(I) chains, but pro-α2(I) chains cannot form homotrimers, as shown in insect cells [14]. Consequently, the trimerization event is not only essential for the initial encounter of the triple helix but also prevents formation of misaligned triple helices [15]. Both the mechanism for correct α-chain selection and the mechanism determining the ratio of α1 and α2 chains defining the collagen composition remain unknown. More is known on the molecular domain responsible for chain selection and triple helix initiation. Upon alignment of the appropriate pro-α-chains, trimerization is initiated through the non-collagenous domain at the C-termini. The globular non-collagenous C-propeptides contain one N-glycan residue and eight cysteine (Cys) residues for α1(I) and seven for α2(I) [16]. The four carboxyl-terminal Cys residues are involved in intramolecular disulfide bonds while the other Cys residues at the N-terminal end

of the C-propeptide form intermolecular disulfide bonds [16, 17] Studies on procollagen III revealed that the non-collagenous C-propeptides align and recognize the three identical α-chains [18, 19]. After bringing the non-collagenous domains in close proximity by forming non-covalent interactions, disulfide linkages are built between the three non-collagenous domains. The newly formed C-terminal trimerization domain serves as a helix nucleation site and propagates triple helix formation in a zipper-like fashion from the C-terminus to the N-terminus. Studies from OI patients harboring distinct Gly to serine (Ser) mutations in the collagenous domain emphasize the C-to-N terminal triple helix formation. The clinical severity of these mutations concords with the distance of the Gly to Ser substitutions from the C-termini, as demonstrated by the lethality associated with mutations close to the C-termini and by the moderate diseases associated with mutations closer to the N-termini of the α-chain [20]. Also mutations in the trimerization domain of collagen type I (*COL1A1*) lead to OI with mild to severe phenotypes. Mostly, these mutations cause a slow assembly of the trimeric suprastructure and secretion of hypermodified collagen type I, or in more severe cases to early stop signals [21-24]. The variable severity of mutations suggests that the carboxyl-terminal propeptide is responsible for the correct interaction and selection of the three pro-α-chains. Similar findings have been reported for mutations in the carboxyl-terminal propeptide of collagen type III (*COL3A1*) and collagen type V (*COL5A1* and *COL5A2*) that lead to Ehlers-Danlos Syndrome (EDS) [25-27].

Peptidyl prolyl cis-trans isomerases like CyPB and FKBP65 might also influence the kinetic of collagen folding [28]. These isomerases mainly locate to the C-terminus where collagen folding is initiated and facilitate triple helix formation. The C-nucleation domain contains four consecutive Gly-Pro-Hyp triplets, of which the pyrrolidines must be in trans configuration in order to be competent to initiate triple helix formation [29]. Additionally, several chaperones, such as GRP78, HSP-47, and FKBP65, are in contact with the collagen molecule (Figure 2) [30]. Collagen-bound HSP-47 might prevent the aggregation of procollagen in the ER and is not released until traveling to the *cis*-Golgi network where it dissociates due to lower pH and returns to the ER [30-32]. The newly formed triple helical procollagen molecules ready for secretion are large in size with a length of 300 – 400 nm. Since secretory vesicles are generally only 60 – 80 nm in diameter, collagen molecules either use a different secretory pathway or have to increase the vesicle size in order to accommodate the large procollagen fibers. The latter has recently been described as the mechanism how collagen molecules ensure their packaging and hence their secretion. The ER luminal procollagen molecules recruit the cytosolic ubiquitin ligase CUL3-KLHL12 possibly via the transmembrane protein TANGO1 to the ER exit site [33, 34]. CUL3-KLHL12 attaches a single ubiquitin to COPII, in particular SEC31 which is a component of COPII. COPII monoubiquitination

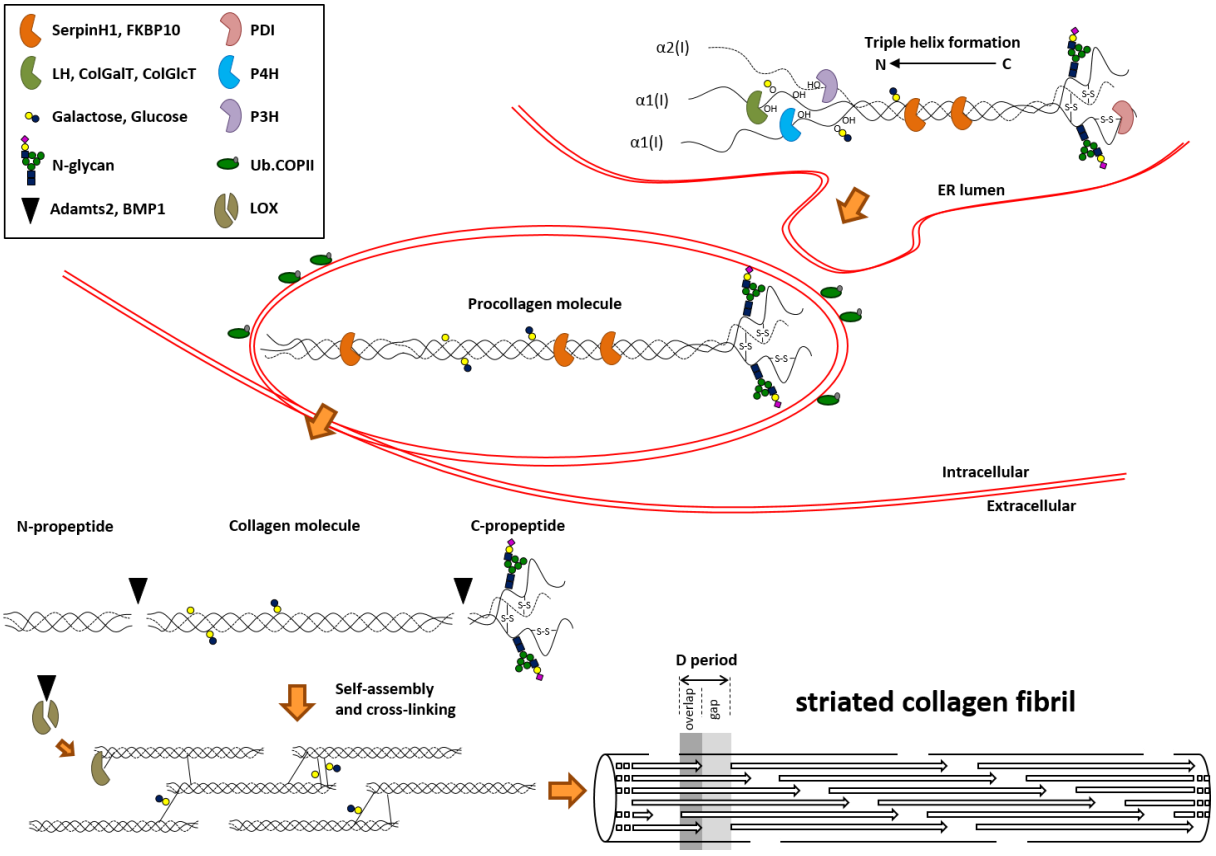


Figure 2| **Biosynthesis of fibril-forming collagen type I.** Upon α-chain synthesis into the rough ER, the chains get hydroxylated (P4H, LH, P3H) and glycosylated (ColGalT, ColGlcT). Two α1(I) chains and one α2(I) chain trimerize and triple helix formation is initiated from the disulfide cross-linked (PDI) C-terminus. The chaperone protected (SerpinH1, FKBP10) triple helical procollagen molecule passes the secretory pathway via enlarged monoubiquinated COPII vesicles. Extracellular proteases cleave off the N-propeptide (Adamts2) and non-collagenous C-propeptide (BMP1) and activate (BMP1) the cross-linking enzyme lysyl oxidase (LOX). The released collagenous domain self-assembles into fibrils and maturates its cross-links to rigid covalent structures. The process of self-assembly is highly defined resulting in a characteristic striated pattern with repeating periodicity (D-period).

drives the assembly of large COPII coated secretion vesicles thereby providing space for the large procollagen molecules [35]. Upon efficient secretion the procollagen molecule gets cleaved at specific non-collagenous sites by extracellular proteases ADAMTS2 and BMP1/tolloid-like proteases releasing the N-propeptide and the globular C-propeptide, respectively. The remaining fibers then self-assemble into fibrils. In particular the cleavage of the C-propeptide is essential for the initiation of self-assembly and fribrillogenesis [36, 37]. BMP1 also cleaves the catalytically quiescent lysyl oxidase enzyme and produces an active mature form (LOX) [37]. LOX catalyzes the oxidative deamination of specific telopeptidyl lysines (Lys) and hydroxylysines (Hyl). The resulting aldehydic forms Lys^{Ald} and Hyl^{Ald} initiate a series of condensation reactions to form covalent intermolecular cross-links with Lys or Hyl residues from adjacent fibers. Depending on the residues involved several pathways yield mature pyridinoline (Pyr), deoxypyridinoline (dPyr), pyrrole (Prl), and deoxypyrrole (dPrl) cross-links (Figure 3) [37].

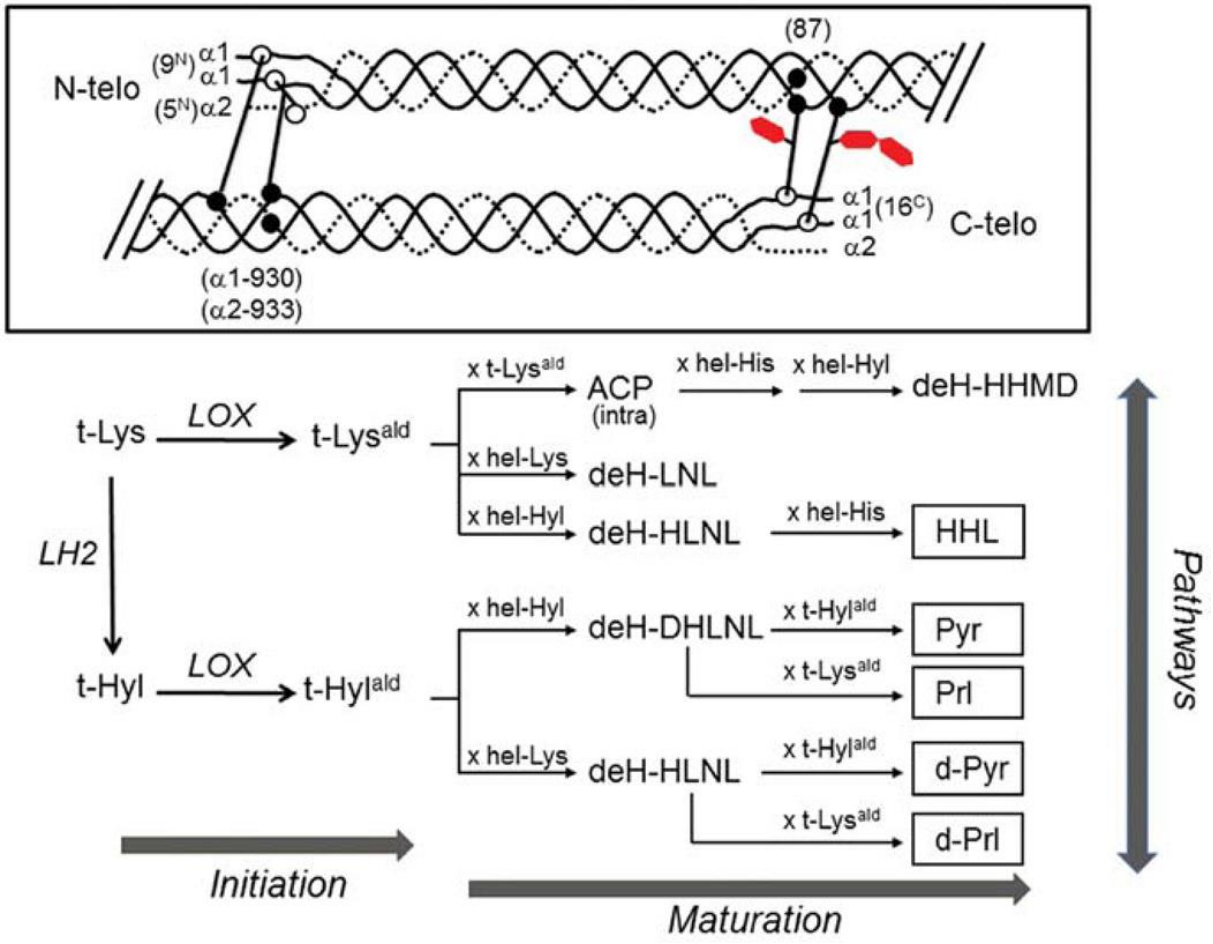


Figure 3| **LOX-mediated fibrillar type collagen cross-linking.** Top panel shows the cross-linking sites with telopeptidyl aldehydes (open circles) t-Lys^{ald} and t-Hyl^{ald}, and the helical Lys and Hyl residues (closed circles) hel-Lys and hel-Hyl, respectively. Red hexagon represent the carbohydrates Gal and Glc attached to the hel-Hyl involved in cross-links. The lower panel summarizes the initiation and maturation of cross-link pathways. The boxed compounds are non-reducible cross-links. Taken from [37].

COLLAGEN PTMS AND THEIR DIVERSITY AT SINGLE LOCI

The intracellular PTMs of collagen encompass five distinct enzymatic reactions which all take place in the ER, namely, 4-hydroxylation of Pro, 3-hydroxylation of Pro, 5-hydroxylation of Lys, galactosylation of Hyl, and glucosylation of galactosylhydroxylysine (Figure 4). Although these modifications were discovered eight decades ago and the structural and enzymatic basis of the modifications is largely solved, the regulation of the PTMs at a given locus is not defined. For example, a single Lys residue can either remain unmodified, be hydroxylated, be hydroxylated and galactosylated, or be hydroxylated, galactosylated, and glucosylated. Such variations at a single locus appear tissue specific but sometimes even vary within the same type of collagen. Profound knowledge of how PTM localization is regulated might shed light on the possible functional contribution of 3-Hyp or glycosylation since no function could be assigned which generally describes the necessity of these types of modification. Knowing the functional properties of PTMs

and their regulatory mechanism might provide new possibilities for therapeutic approaches for many hereditary diseases. Additionally, a complete picture of the PTM distribution along the individual collagen types is desirable for the production of collagen for biotechnological purposes.

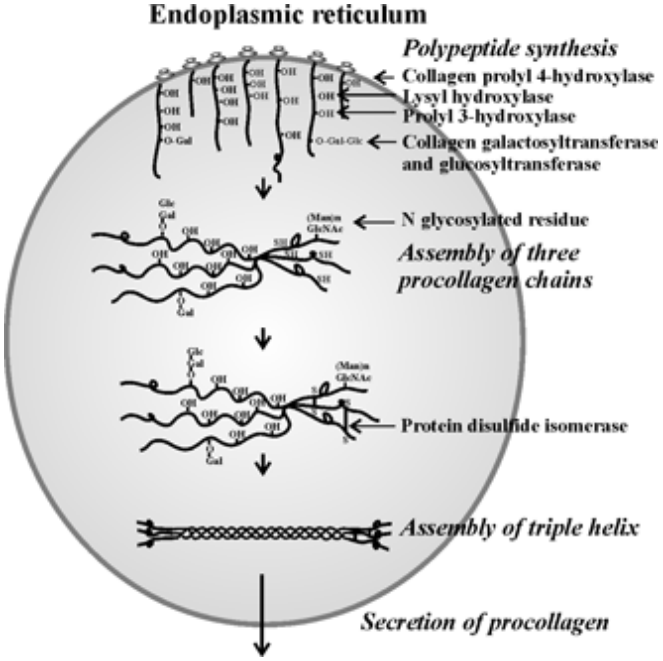


Figure 4| **Intracellular post-translational modifications of collagen.** Upon polypeptide synthesis into the rough ER, the α-chain gets modified by prolyl 4-hydroxylase, prolyl 3-hydroxylase, and lysyl hydroxylase. The latter modification forms hydroxylysine which get O-glycosylated by galactosyltransferase and glucosyltransferase. The non-collagenous globular domain at the C-terminus get N-glycosylated and disulfide linked prior to assembly into triple helix and secretion to the extracellular space. Taken from [38].

Proline hydroxylation

Hydroxylation of Pro residues is the most abundant PTM in collagen. The Pro residues either get hydroxylated at the C4, which is most common, or at the C3 position giving 4-Hyp or 3-Hyp, respectively. In mammalian fibrillar collagen α-chains, about a quarter of the amino acids are prolyl residues with about half of them being 4-hydroxylated. 4-Hyp is mainly found at the Yaa position of the common Gly-Xaa-Yaa motif. Since the collagenous domain of type I fibrillar proteins encompass about 1'000 amino acids, it is common to present a single amino acid frequency as per 1'000 amino acids. Mass spectrometric mapping of hydroxyproline in collagen type V revealed 106 4-Hyp / 1'000 amino acids and 98 4-Hyp / 1'000 amino acids for bovine and recombinant human α1(V)-chain, respectively [39]. The sites of 86 Pro residues were found hydroxylated in both chains suggesting that the PTM at these sites is invariant and always 4-Hyp.

The conversion of Pro to 3-Hyp is a rare but conserved modification of many collagens. The 3-Hyp content is variable among collagen types with an occurrence of one to two residues per α-chain in types I and II, between three to six residues in types V and XI and more than 10 residues in type

IV [40-42]. Unlike 4-Hyp, 3-Hyp is thought to be exclusively found in the Xaa-positioned Pro and restricted to Gly-3Hyp-4Hyp triplets. However, recent studies mapping Hyp based on mass spectrometric techniques unraveled new sites for Xaa position hydroxylation besides Gly-Hyp-Hyp. Xaa-positioned Hyp has been identified in Gly-Hyp-Ala, Gly-Hyp-Val, and Gly-Hyp-Gln triplets [39, 43]. Due to the inability to distinguish 3-Hyp from 4-Hyp with conventional mass spectrometry, it is not clear whether the assigned Hyp is 3-hydroxylated or 4-hydroxylated. Hence, the principle that every Xaa-positioned Hyp is 3-hydroxylated and every Yaa-positioned Hyp is 4-hydroxylated is doubted. Since amino acid analysis of type II collagen reveals only one to two 3-Hyp residues per α-chain, the mass spectrometric identification of 14 Xaa-positioned Hyp residues per α-chain might indicate that also 4-Hyp exists at the Xaa position [43]. Xaa-positioned 4-Hyp residues are thought to destabilize the triple helix in Gly-4Hyp-Pro triplets [44] but not in the context of triplets that lack Pro or Hyp at the Yaa position such as Gly-Hyp-Ala triplets [45]. Otherwise, the data from mass spectrometric Hyp mapping could be interpreted differently. Assuming that not every 3-Hyp site per α-chain is 100% hydroxylated, the one to two 3-Hyp per type II α-chain might be localized to variable sites within the chain. Such an interpretation would be in accordance with the amount of 3-Hyp per chain and the amount of possible 3-Hyp sites, however the regulating mechanism defining the localization is unknown.

Lysine hydroxylation and glycosylation

Lys is hydroxylated at the C5- (or epsilon) position forming Hyl. A distinct type of O-glycosylation characteristic for collagen is extending on these Hyl residues. First, a galactose (Gal) residue is linked via a β1-O glycosidic bond to the peptidyl 5-hydroxylysine forming the monosaccharide structure Gal-Hyl. Sometimes the monosaccharide is elongated with a single glucose (Glc) residue to form the disaccharide structure Glc-Gal-Hyl (Figure 5).

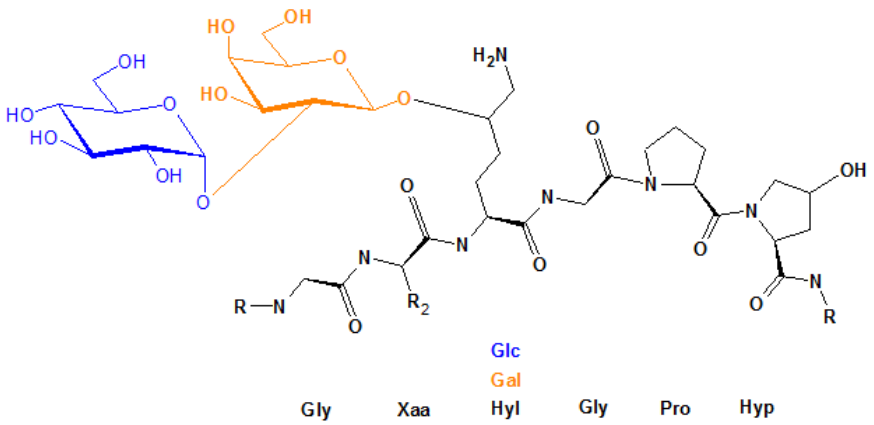


Figure 5| **Dissaccharide structure attached to Hyl of the collagen backbone.** Gal (orange) is attached via β1-O glycosidic bond to the C5-hydroxy group of Hyl. Glc (blue) is then linked via α1-2 glycosidic bond to the galactosyl-hydroxylysine polypeptide chain.

In mammalian collagens the carbohydrate content of hexoses ranges from 0.4% in skin to about 4% in cartilage and 12% in basement membrane [13]. This goes in hand with the predominant types of collagen found in these tissues, namely type I in skin, type II in cartilage, and type IV in basement membranes separating epithelial and endothelial tissues from underlining connective tissues (Figure 6). Dermal fibrillar type collagen carries fewer glycosylated residues than network forming basement membrane type IV collagen which has a general high degree of PTMs. In this context, the type of collagen presumably dictates the amount of modification needed to fulfill a given function in the respective tissue. However, whether the function of the modified collagen remains the same irrespective of the tissue where the collagen is expressed, remains to be investigated. As the amount and ratio of the disaccharide to the monosaccharide vary substantially between different collagen types, the two saccharides might have different functions from each other (Figure 6). Most of the interstitial collagen types have a molar ratio of 2:1 for Gal:Glc residues meaning as much disaccharides as monosaccharides. The basement membrane type IV collagens carry more disaccharides than monosaccharides on Hyl [46]. The overall amount of glycosylation in a single chain might depend on the helix folding kinetics but the mechanisms determining the localization and regulating the extent of glycosylation are not known. Unexplained is also the finding that the site of modifications in a given collagen type vary even though they are expressed in the same tissue. This suggests that the overall amount of modified residues might be more important than a specific site within the molecule. In the α1-chain of type V collagen the lysine residue 84 has been detected as Gal-Hyl form and Glc-Gal-Hyl form suggesting that glycosylation at this site is dynamic [39]. Comprehensive mapping of glycosylation

Residues / 1000 amino acids					
	Type-I	Type-II	Type-III	Type-IV	Type-V
Hyl	5-8	23	5	45	23-43
Gal-Hyl	1	4	0.5	2	3-5
Glc-Gal-Hyl	0.3	12	3	32	20-29

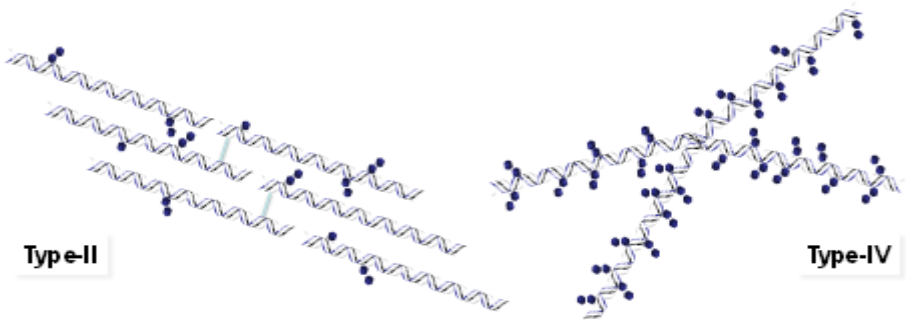


Figure 6| **Variability of collagen modification.** The distribution of Lys hydroxylation and glycosylation vary significantly between different types of collagen. In general, classical fibrillar types (I,II,III) carry less modifications than network-forming mesh-like types (IV). Short chain fibrillar type V with retaining N-propeptides are also highly glycosylated. Taken from [46].

sites in collagen type II supports the findings for dynamic glycosylation at specific sites [43]. The site mapping analysis also revealed that glycosylation not only depends on hydroxylation of Lys residues but also other factors such as steric hindrance from adjacent amino acids. Recently, the distribution of lysine hydroxylation and glycosylation at a single locus has been investigated (Figure 7). The results indicated that the variability does not only depend on the type of tissue but also on the species [47].

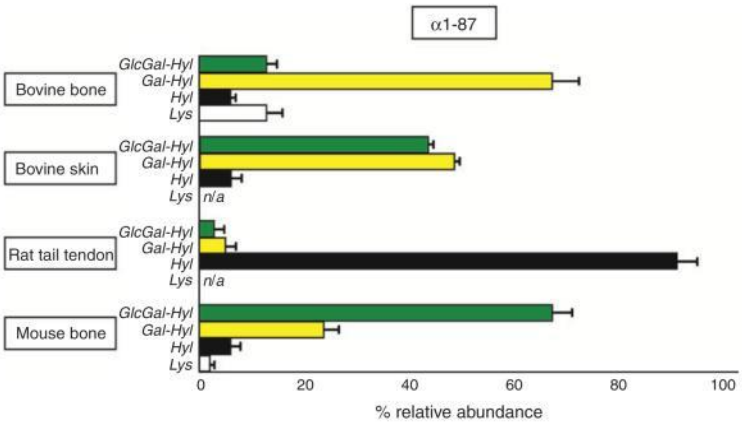


Figure 7| **Modification variability of a single locus among tissues and species.** Distribution of Lys, Hyl, Gal–Hyl and Glc-Gal–Hyl at residue Lys^{87} of the α1-chain in bovine bone, bovine skin, rat-tail tendon, and mouse bone, determined by a semi quantitative mass spectrometric approach. Taken from [47].

FUNCTIONS OF COLLAGEN PTMS

The PTMs of collagen significantly contribute to collagen's function providing the connective tissue with strength and shape. The importance of the PTMs can be exemplified by the disease scurvy, the old sailor's illness, caused by vitamin C (ascorbate) deficiency. People deficient for ascorbate show symptoms such as easy bruising, fragile capillaries, poor wound healing, skin changes and bone pain. Ascorbate is an important cofactor for the Pro and Lys hydroxylase enzymes. Hydroxylation of Pro and Lys residues in procollagen α-chains is reduced upon ascorbate deficiency in the cells. The resulting defect in collagen biosynthesis leads to collagen malfunctioning and unstable collagen fibers as observed in scurvy [48].

To date genetic defects in the human P4H complex leading to the absence of 4-Hyp have not been reported. The absence for P4H deficiencies in humans could be due to lethal consequences at early embryonic stages. Lacking 4-Hyp leads to failure of collagen type IV assembly and to impaired collagen secretion. Consequently, decreased deposition of collagen in the extracellular space results in dysfunctional basement membrane integrity. Supporting evidence comes from studies done in the nematode worm *C. elegans*. The P4H complex is essential for viability and

morphogenesis in *C. elegans*. RNA-interference studies revealed that disruption of the genes encoding the P4H complex resulted in embryonic lethality [49, 50]. A single study from a *P4h1* null mice which die at embryonic day 10 emphasizes the lethality of an absent P4H complex and thus suggesting defective architecture of basement membrane to be the main cause of death in these mice [51].

4-hydroxyproline and helix thermal stability

The contribution of 4-Hyp to thermal collagen stability is established [1], however the mechanism how 4-Hyp modification stabilizes the helical structure is not fully clear. Water-bridged hydrogen bonding between 4-Hyp residues and Gly residues are thought to be essential for the firm attachment of each adjacent α-chains [52]. On the contrary, the contribution of water-bridges to helix stability is in so far unlikely to be significant since the entropic cost to immobilize more than 500 water molecules per triple helix is enormous in order to build and maintain the bridges [53]. Additionally, $(Gly\text{-}Pro\text{-}Hyp)_{10}$ peptides maintain a stable triple helix even in anhydrous environments like methanol or propan-1,2-diol [54]. Later studies identified a stereo electronic effect by which the electron withdrawing oxygen from 4-Hyp pre-organizes the chain in a proper conformation for triple helix formation [53, 55, 56]. Studies using the more electronegative fluorine atom in a fluoro-proline (Flp) conjugation, which cannot participate in potential hydrogen bonding networks, substantiate the stereo electronic effect. Flp in the $(Gly\text{-}Pro\text{-}Flp)_{10}$ peptides accomplish more stable helix formations than $(Gly\text{-}Pro\text{-}Hyp)_{10}$ peptides [53].

3-hydroxyproline and type specific functions

With the identification of three 3-Hyp residues in collagen type V, each with approximately 234 residues apart from each other, the theory of a highly determined distribution pattern of 3-Hyp residues among fibrillar type collagens arose [40]. The 234 residues distance resembles the D-period length characteristic in fibrillar type collagens. Since this periodicity of 3-Hyp appearances has been detected mainly for collagen types V and XI which are minor components of type I and II fibrils, respectively, it is possible that 3-Hyp sites facilitate the staggered D-periodic alignment. Types V and XI might serve as a template for proper aligning and self-assembly of the major collagen types I and II via hydrogen bonding between adjacent 3-Hyp residues in a staggered array. The invariant 3-Hyp site at Pro^{981} which has been found in types I, II, and V might provide the counterpart for hydrogen bonds with 3-Hyp sites of adjacent chains. Hence, the D-periodicity of 3-Hyp residues could be proposed to possess a fundamental role in ordered self-assembly of the fibrillar type supramolecular structure [40]. In contrast, the non-fibrillar but network forming type IV collagen contains a high amount of 3-Hyp residues. Whether 3-Hyp plays a role in ordered assembly for this type as well is unknown. The large amount of 3-Hyp in type IV collagen might have a different function than in fibrillar types. The 3-hydroxy pyrrolidine ring from 3-Hyp is

exposed at the surface of the triple helix thereby contributing to epitopes that underlie immune responses against collagens. Sub endothelial type IV collagen bears the platelet-specific glycoprotein VI GPVI-binding site which might be 3-Hyp. Exposure of 3-Hyp due to a damage in the basement membrane initiates platelet aggregation and leads to blood coagulation upon injury [57]. That mechanism provides the link between blood coagulation and the extracellular matrix (ECM).

5-hydroxylysine and fibril cross-linking

Hyl basically accomplishes two distinct functions. On the one hand, the more distal Hyl in the N- and C-telopeptidyl regions of collagen can be oxidatively deaminated to produce reactive aldehydes. These aldehydes undergo a series of non-enzymatic condensation reactions with Lys or Hyl from adjacent procollagen fibers. Lys, Hyl, and their aldehydic forms are necessary for intra and intermolecular collagen cross-link formation (Figure 3) (reviewed in [37]). On the other hand, Hyl, particularly located in the helical domain, serves as acceptor site for carbohydrate attachments. Hydroxylation and subsequent glycosylation of Lys in the Gly-Xaa-Lys motif influences many biological functions, including fibrillogenesis [58], crosslinking [59, 60], and matrix mineralization [61].

Glycosylated hydroxylysine and modulating cell receptor binding

Alterations of collagen glycosylation have often been reported in several bone and skeletal disorders thereby suggesting that collagen glycosylation might play a role in bone mineralization [61-63]. The glycan might regulate the distribution of bone mineral along the collagen fibril. However, the localization of the glycan to Hyl involved in crosslinking anticipates the functional involvement of the glycan in cross link formation. During cross link maturation the glycan might regulate the cross link species towards either divalent or trivalent cross links depending whether the involved glycan is in the mono- or disaccharide form [60]. Such a mechanism would also explain the variable extent of Hyl glycosylation. Other functions for collagen glycosylation might target the collagen remodeling process. The endocytic collagen receptor μPARAP/Endo180 internalizes collagen for lysosomal degradation via its fibronectin II domain. Additionally, the receptor carries a lectin domain which has been shown to modulate the endocytic efficiency towards highly glycosylated type IV collagen [64]. However, the impact of the glycan is uncertain, since another endocytic receptor, the mannose receptor, does not share this property as the receptor internalizes glycosylated collagen independent of a functional lectin domain [64]. The potential for the glycan to serve as receptor is also described for interactions of cells with basement membranes. The high glycan content in type IV collagen could serve as interaction or binding receptor in order to recruit cells to the basement membrane. Glycosylated Hyl residues thereby modulate cell adhesion through integrin binding [65]. In contrast, the glycan might have

different functions in fibrillar type collagens than the basement membrane type IV collagen. In OI patients, highly glycosylated collagen fibrils show a slight increase in fibril diameter, however it is not clear whether the increased glycosylation or the absence of 3-Hyp due to defective enzymatic activity in these patients is responsible for the disturbance of the lateral fibril growth [66].

In summary, several function for collagen glycosylation including control of matrix mineralization, crosslinking, collagen remodeling, collagen-cell interaction, and fibrillogenesis have been reported. Despite all the reported findings explaining the function of collagen glycosylation, the specific biological function of glycosylated hydroxylysine in relation to the extent and type of glycosylation at their molecular loci, is still not clearly defined.

COLLAGEN MODIFYING ENZYMES

The collagen Pro and Lys hydroxylating enzymes belong to the protein family of 2-oxoglutarate dioxygenases and require iron (Fe^{II}), O_2, 2-oxoglutarate and ascorbate for their activity. Molecular oxygen is captured with iron (Fe^{II}) in the enzyme's active pocket to hydroxylate Pro or Lys. One oxygen atom of O_2 makes up the hydroxyl group on Pro or Lys, whereas the other atom is incorporated into succinate. Succinate and CO_2 are produced in a coupled reaction by oxidative decarboxylation of 2-oxoglutarate (Figure 8). Upon oxidizing the iron Fe^{II} to Fe^{III}, the enzyme becomes inert until ascorbate reduces the enzyme's iron, restoring its active state. The availability of ascorbate in that reaction is significant, since a lack of ascorbate leads to reduced hydroxylation and unstable collagen fibers as observed in scurvy [67].

Prolyl hydroxylases

The mammalian enzyme prolyl 4-hydroxylase (P4H) is a $\alpha_2\beta_2$ tetramer that catalyzes the formation of 4-Hyp. The α subunit P4Ha contains the substrate-binding domain and the catalytic active site. In humans, three genes code for isoforms of the α subunit, namely *P4HA1*, *P4HA2*, and *P4HA3.* The β subunit P4Hb functions independently as a protein disulfide isomerase (PDI) and is encoded by *P4HB* [68, 69]. P4Hb binds and retains the α subunit in the ER through the C-terminal ER-retention signal KDEL and maintains the α subunit in a soluble active form [70]. The hydroxylation reaction is only performed on individual collagen strands but not on triple helices [71] and P4H does not hydroxylate single Pro residues but recognizes the minimal Gly-Xaa-Pro motif [72].

Like P4H, the prolyl 3-hydroxylase (P3H) also belongs to the 2-oxoglutarate dioxygenase protein family and requires Fe^{II}, O_2, 2-oxoglutarate and ascorbate as cofactors (Figure 8). In humans, three homologues sequences have been identified and designated as P3H1, P3H2 and P3H3 encoded by

leprecan, *MLAT4* and *LEPREL2*, respectively [73]. P3H1/Lepre1 was first described as a basement membrane-associated leucine- and proline-enriched proteoglycan [74] and shares 46 and 41% sequence identity to P3H2 and P3H3, respectively. Unlike the $\alpha_2\beta_2$ tetrameric P4H complex, immune-purified P3H is able to hydroxylate procollagen substrates by itself [73]. Still, P3H forms a tight complex with cyclophilin B, a peptidyl prolyl cis-trans isomerase (CyPB) and with the chaperone cartilage-associated protein (CRTAP). The importance of CRTAP for prolyl 3-hydroxylation becomes evident since defective CRTAP leads to decreased prolyl 3-hydroxylation [75]. Moreover, the absence of CRTAP and 3-Hyp results in over modified collagen fibrils with increased diameter implying changes in fibrillogenesis. The over modification is due to extended lysyl hydroxylase and P4H activity on the unfolded collagen chain. The proper 3-hydroxylation and positional rotation of the Pro residue at position 986 (Pro^{986}) in α1(I) and α2(I) seems to be important for collagen folding. Mice lacking CRTAP or humans with CRTAP mutations show severe osteochondrodystrophy characterized by severe osteoporosis and higher mineral content of bone matrix [7, 75]. Deficiencies in CRTAP or P3H1 are associated with the clinical features of recessive OI type VII.

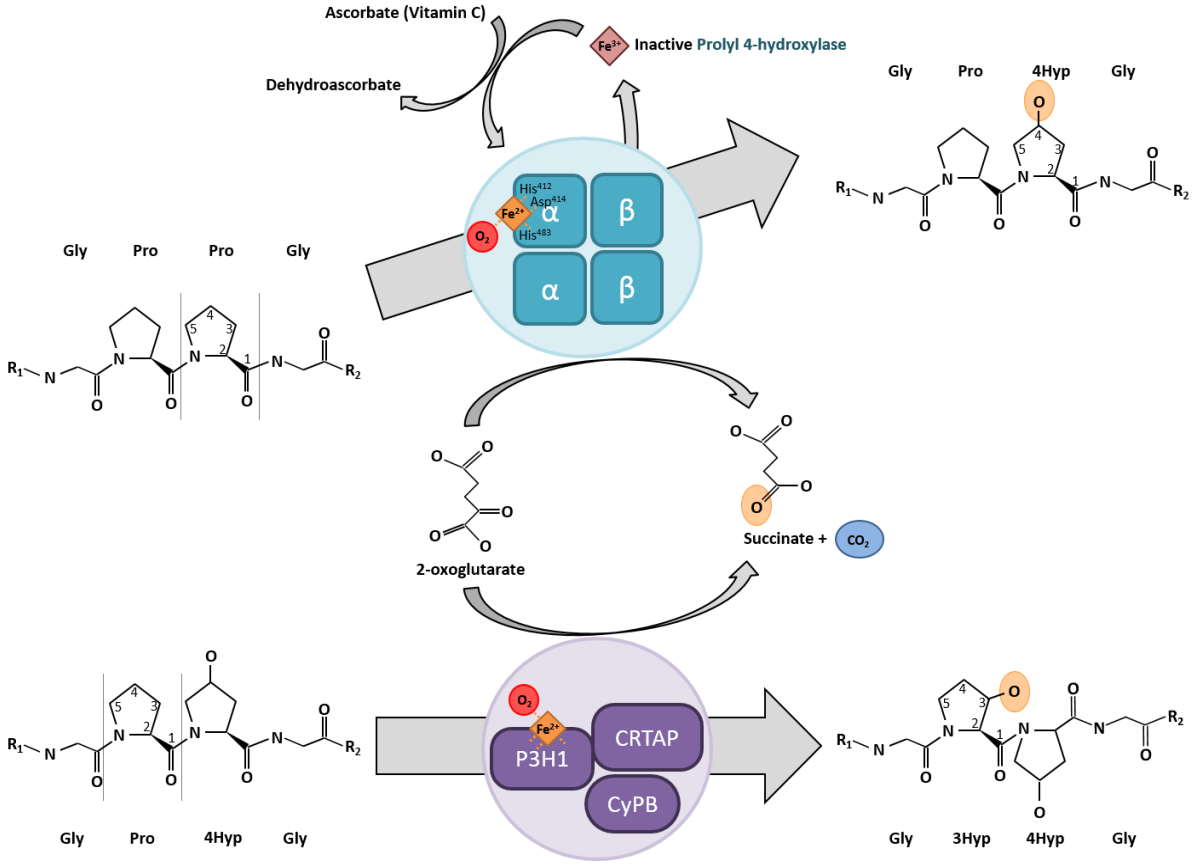


Figure 8| **Hydroxylation of the collagen peptide chain by prolyl 4-hydroxylase (P4H) and prolyl 3-hydroxylase (P3H).** Upper panel: The mammalian P4H is an $\alpha_2\beta_2$ tetrameric enzyme. The α-subunit contains the active site with tightly bound ferrous iron to two histidine (His^{412} and His^{483}) and one aspartic acid (Asp^{414}) residues. By decarboxylating 2-oxoglutarate to succinate, one oxygen atom of O_2 gets incorporated into succinate and the other is used to hydroxylate Pro at the C4-position (orange shaded O). Ascorbate reduces the oxidized ferric iron to ferrous iron, thereby restoring the enzymes active state. Bottom panel: The same mechanism is used by the P3H complex to hydroxylate Pro at the C3-position (orange shaded O).

Lysyl hydroxylases

The collagen lysyl hydroxylases (LH) catalyze hydroxylation of Lys residues by a mechanism similar to that described for prolyl hydroxylases. The addition of a hydroxyl group at the C5-position of a polypeptide-lysine results in the formation of Hyl accomplished by LH. LH enzymes form homodimers and are peripheral membrane proteins in the lumen of the ER [76, 77]. In humans, three isoforms exist, called the procollagen-lysine,2-oxoglutarate 5-dioxygenase 1 (*PLOD1*), *PLOD2* and *PLOD3* encoding for LH1, LH2 and LH3, respectively [9, 78]. Alternative RNA splicing has been observed only for the *PLOD2* gene, resulting in a shorter LH2a and a longer LH2b splicing variant with an extra exon [79]. LH1 and LH2 share 59% identity and LH3 has 57% identity with LH1. Phylogenetic analyses from mouse *Lh* cDNA proposed that the LH1 and LH2 have been brought about by a duplication event, since they are more closely related to each other than to LH3 [80].

The expression pattern of all three isoenzymes is rather widespread than tissue specific. All three or at least two isoenzymes have been found in the same tissues analyzed from human and mouse [79-82]. Conclusively, the isoenzymes might not only appear as tissue specific variants but could have substrate specificity in the same tissue. Different substrates could include the type of collagen or different acceptor sites on the same collagen α-chain. Certainly, the extent of hydroxylysine in collagen type I varies between bone and skin collagen in the telopeptidyl domain but does not vary in the collagenous triple helical domain [83]. That finding indicates the participation of two different isoenzymes in hydroxylating Lys residues on the same collagen type. Indeed, LH1 has been identified to be mainly responsible for Lys hydroxylation in the collagenous triple helical domain with Lys in the Yaa position of the Gly-Xaa-Yaa acceptor motif [84-86]. In contrast, LH2 specifically hydroxylates telopeptidyl Lys residues in the Xaa-Lys-Ala or Xaa-Lys-Ser sequence [83, 87]. The specific substrate for the third isoenzyme LH3 is unknown. However, LH3 has been postulated to have triple enzymatic activity as in hydroxylation of Lys, galactosylation of Hyl and glucosylation of Gal-Hyl [88]. LH3 is a classical lysine hydroxylase with a Fe- dioxygenase domain at the C-terminus. A possible galactosyltransferase activity residing in the N-terminus could never be confirmed unlike the glucosyltransferase activity, which has been shown repeatedly [89, 90]. The contribution of PLOD3 to collagen glucosylation appears to be on a low level compared to collagen galactosyltransferase and it is questionable whether it is of sufficient biological relevance [91].

Glycosyltransferases

Even though glycosylation of collagen was identified in 1935 [92] and the structures of the glycosides have been described since 1966 [5, 93] some of the glycosylating enzymes were discovered only in the past two decades. However, the publication of *PLOD3* being an enzyme with

triple catalytical function namely Lys hydroxylation and subsequent galactosylation and glucosylation led to a lack of interest in continuing with the scientific investigation to identify the collagen glycosylating enzymes. Nevertheless, in 2009 the galactosyltransferase could be identified and annotated as *GLT25D1* and *GLT25D2* [6]. *GLT25D1* is the prevailing isoform with a widespread expression pattern among human tissues unlike *GLT25D2* which seems to be more restricted to the nervous system and the skeletal muscle. GLT25D1 is a soluble endoplasmic reticulum localized protein predefined by the C-terminal RDEL ER-retention signal sequence and verified by immunofluorescence staining [94]. Experiments with lysates from semi-purified glycosyltrasnferases reveal that the collagen glycosylating enzymes depend on divalent metal ions, preferentially Mn^{2+}, and a nucleotide activated hexose donor to fulfill their catalytic activity [5]. The specific collagen glucosyltransferase (ColGlcT, EC 2.4.1.66) has not been cloned to date.

Viral collagen modifying enzymes

The characteristic Gly-Xaa-Yaa collagen and collagen-like structure has long been thought to be restricted to metazoans and some prokaryotes [95, 96]. Little is known about the prokaryotic collagen-like proteins and none of them have been characterized for post-translational modifications necessary for collagen integrity. The description of an aquatic giant virus belonging to *Mimiviridae* [97] has shown that collagen-like genes are also contemporary in the viral kingdom. *Acanthamoeba polyphaga* mimivirus expresses seven collagen genes [98, 99]. In addition to collagen genes, these viruses harbor genes encoding P4H [100] and LH enzymes [101]. Indeed, the *Acanthamoeba polyphaga* mimivirus LH L230 efficiently hydroxylates mimiviral and human collagen-like peptides [101].

Since the virus contains genes on its own for hydroxylation of collagen-like proteins the question arose whether they also glycosylate collagen. Virus mediated glycosylation is a very rare feature for viruses and not long known. First evidence suggesting viruses might code for their own glycosylating genes was described in the *Paramecium bursaria* chlorella virus [102, 103]. Recently, analysis of the reported viral genome of *Acanthamoeba polyphaga* mimivirus revealed that mimivirus contains at least eleven open reading frames coding for possible glycosylating enzymes [97]. Together with the identification of sugar biosynthetic pathways for UDP-L-rhamnose [104], UDP-GlcNAc [105] and unusual amino sugar viosamine [106], it is suggested that *Acanthamoeba polyphaga* mimivirus also encodes its own glycosylation machinery. By now, only one mimiviral glycosyltransferase has been cloned and characterized, the bifunctional collagen LH and core glucosyltransferase L230 [101]. Unlike human collagen glycosylation, the mimiviral collagen core glycosyltransferase adds a Glc residue to Hyl of the collagen backbone. In contrast to the human core Gal-Hyl monosaccharide which can be elongated to the Glc-Gal-Hyl

disaccharide structure, it is not known whether the mimiviral core Glc-Hyl structure can be elongated.

Interestingly, the viral collagen modifying enzymes are soluble and active when expressed in E.*coli* [101]. In contrast, mammalian enzymes are not active in bacterial expression systems and the prolyl hydroxylases need co-expression of the partnering subunits, like the PDI, to be active in yeast or tobacco expression systems [107, 108]. The identification of these viral collagen modifying enzymes and their capability to be functionally expressed in bacteria is very interesting for biotechnological purposes thus opening new possibilities for the high yield production of recombinant hydroxylated collagen in bacterial expression systems.

DEFECTS IN COLLAGEN MODIFYING ENZYMES

Heritable disorders of connective tissues including skin, bone, cartilage, blood vessels and basement membranes are among the most common human genetic diseases. The diseases are classified according to clinical features and the pattern of inheritance into EDS, OI, Chondrodysplasias, Alport syndrome, Epidermolysis bullosa and the Marfan syndrome. The classification of these diseases cannot only be made by the link from genotype to phenotype. Defects in collagen type I can lead to OI or EDS. Even more, the same mutation can develop different phenotypes with respect to severity of the disease. For many of these diseases the molecular cause lays in mutations of different collagen types, collagen modifying enzymes or collagen interacting proteins. Until now, defects in eight collagen modifying enzymes have been identified in humans (Table 1). In this chapter the two most prominent diseases of collagen modifying enzymes EDS and OI will be described.

Table 1: Characteristics of collagen modifying enzymes and their association to human disorders.

Gene	Enzymatic function	Diseases caused by gene mutation [Reference]
Prolyl 4-hydroxylase and protein disulfide isomerase		
P4HA1	Hydroxylates Pro at the C4 position	Embryonic lethal in mice [51]
P4HA2	Hydroxylates Pro at the C4 position	
P4HA3	Hydroxylates Pro at the C4 position	
P4HB	Protein disulfide isomerase	Essential for viability in *C.elegans* [109]
PDIA3	Protein disulfide isomerase	
PDIA4	Protein disulfide isomerase	
PDIA6	Protein disulfide isomerase	
Prolyl 3-hydroxylase and peptidyl prolyl cis-trans isomerase		
LEPRE1	Hydroxylates Pro at the C3 position	OI type VIII [8, 110]
P3H2	Hydroxylates Pro in Col. type IV	High myopia [111]
P3H3	Hydroxylates Pro at the C3 position	
CRTAP	Cofactor for P3H1 (*LEPRE1*)	OI type VII [75, 112]
PPIB	Peptidyl prolyl cis- trans isomerase	OI type IX [113]
FKBP10	Peptidyl prolyl cis- trans isomerase	OI type XI [114]
Lysyl hydroxylase		
PLOD1	Hydroxylates helical Lys	EDS type VI [86]
PLOD2a		
PLOD2b	Hydroxylates telopeptidyl Lys	Bruck syndrome 2 [115]
PLOD3	Hydroxylates Lys in col. type IV, V	
Glycosyltransferase		
GLT25D1	Galactosylates Hyl	
GLT25D2	Galactosylates Hyl	
EC 2.4.1.66	Glucosylates Gal-Hyl	
Lysyl oxidase		
LOX	Oxidatively deaminates Lys/ Hyl	Lathyrism [116]
Metalloproteinase		
ADAMTS-2	Cleaves the N-terminal propeptides	EDS type VIIc [117]
BMP-1	Cleaves the C-terminal propeptides	OI type XII [118, 119]
Tolloid-like 1	Cleaves the C-terminal propeptides	

Osteogenesis imperfecta

OI is a heritable bone dysplasia characterized by fragile bones, easy susceptibility to fractures and growth deficiency [120]. OI defines a heterogeneous group of diseases with variable severity and occurs in approximately 15'000 to 20'000 births [121]. Most cases affect mutations in collagen type I with autosomal dominant inheritance. Only 10% mutant gene expression is sufficient for disruption of normal collagen function [122]. Several rare cases have been identified with mutations in non-collagenous genes. But the function of all of these genes relates to collagen biosynthesis. Currently, the genetic classification of OI distinguishes between autosomal dominant and autosomal recessive inheritance but with a broad range of severity across both groups. The dominant types encompass the classical types caused by collagen type I mutations and the recessive types group defects of genes involved in collagen modification, folding and processing (Table 2).

Table 2: Classification of Osteogenesis imperfecta. From [120, 121, 123].

OI Type	Gene Defect	Phenotype
Dominant inheritance		
Classical types		
I	*COL1A1* null allele	Mild, nondeforming
II	*COL1A1* or *COL1A2*	Perinatal lethal
III	*COL1A1* or *COL1A2*	Progressively deforming
IV	*COL1A1* or *COL1A2*	Moderately deforming
COL1-mutation negative		
V	*IFITM5*	Distinct histology
Recessive inheritance		
Mineralization defect		
VI	*SERPINF1*	Distinct histology
3-Hydroxylation defects		
VII	*CRTAP*	Severe to lethal
VIII	*LEPRE1*	Severe to lethal
IX	*PPIB*	Moderate to lethal
Chaperone defects		
X	*SERPINH1*	Severe
XI	*FKBP10*	Progressive deforming, Bruck syndrome 1
C-Propeptide cleavage defect		
XII	*BMP1*	Severe, high bone mass case
Unclassified Osteogenesis imperfecta-like		
Zinc-finger transcription factor defect	*SP7*	Moderate
Cation channel defect	*TMEM38B*	Moderate to severe
WNT signaling pathway defect	*WNT1*	Moderate, progressively deforming
Bruck syndrome 2	*PLOD2*	Joint contractures

Autosomal dominant Osteogenesis imperfecta

A general decrease in bone mineral density (osteopenia) and brittle bones are the characteristic clinical features of dominant OI. Beside bone fragility, many patients show signs of dental abnormalities, progressive hearing loss and blue sclera, which are caused by thinness of collagen layers [124]. About 90% of OI cases have heterozygous mutations in either of the collagen type I pro-α chains (*COL1A1, COL1A2*) [125]. Despite the high causative percentage for a single protein, OI patients show a large genotypic heterogeneity with respect to mutation sites. Unrelated patients rarely carry the same mutation in the same gene and even phenotypic variability has been reported for both, related or unrelated patients with the exact same collagen mutation [121, 126]. About 80% of collagen mutations are Gly substitutions and about 20% alter splicing sites [126]. The substitution of Gly with bulkier amino acids leads to assembly of collagen fibers that are branched or abnormally thick and short (Figure 9). Splice site mutations often lead to frameshifts that produce structurally abnormal but partially functional collagen resulting in a mild form of OI. General, mutations in the C-terminal domain are more severe, as these mutations prevent initial triple helix formation which results in a process called collagen suicide and the degradation of the

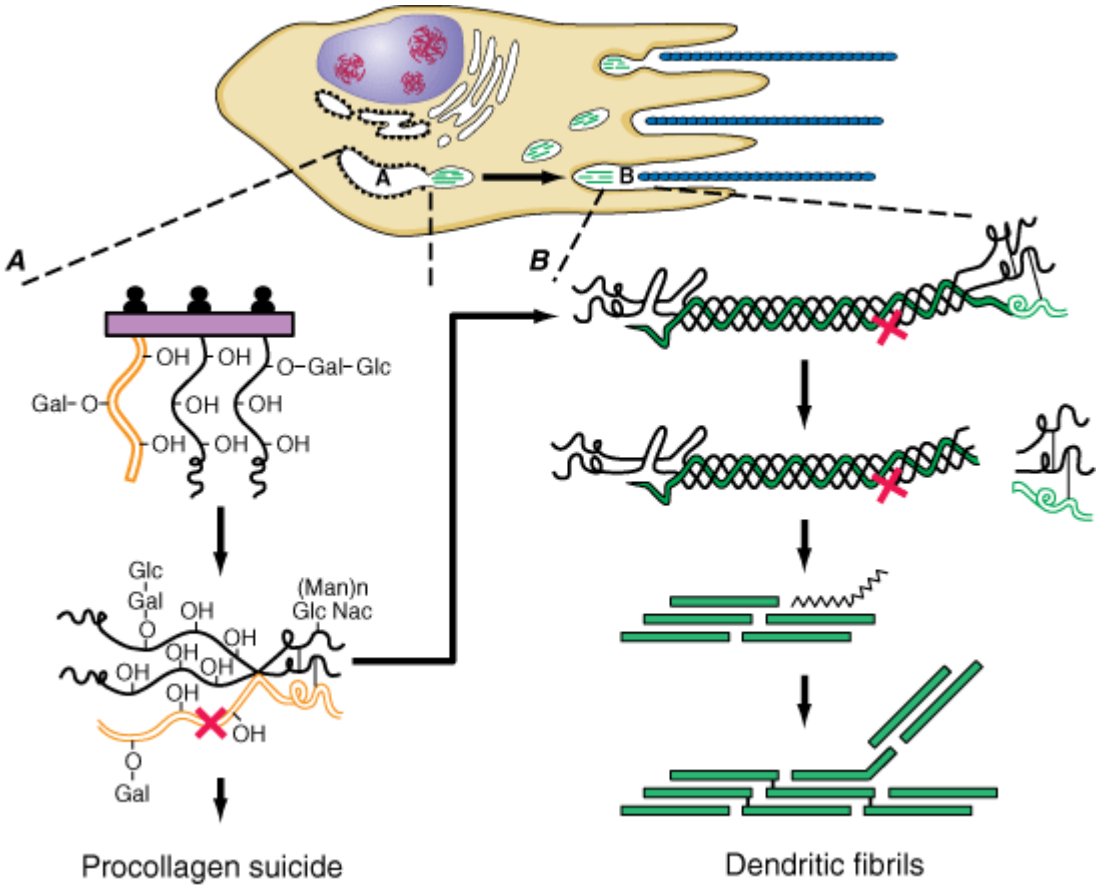


Figure 9| **Defective collagen biosynthesis.** Mutations in one of the collagen pro-α(I) chains can lead to early stop in α-chain synthesis or disability to fold into the triple helix leading to chain degradation and subsequent procollagen suicide. Some mutations sill lead to partially functional but abnormal collagen secretion with branched fibers. (Source: Fauci et al., *Harrison's Principles of Internal Medicine*, 17th Edition: http://www.accessmedicine.com)

whole collagen molecule (Figure 9). Recently, several degrading pathways have been proposed for defective collagen disassembly including ERAD proteasomal degradation of chains, an unidentified pathway for heterotrimeric structures and autophagy of supramolecular aggregates [121]. Some severe to lethal mutations are located to C-terminal regions important for binding of integrin, matrix metalloproteinases, fibronectin and cartilage oligomeric matrix protein (COMP) suggesting erroneous interactions with the extracellular matrix [126].

Autosomal recessive Osteogenesis imperfecta

Other genetic causes, besides mutations in collagen, currently account for about 2 – 5% of OI cases. These non-collagenous genes are mainly inherited in an autosomal recessive pattern and encompass genes coding for collagen modifying enzymes, collagen chaperoning genes and genes involved in osteoblast maturation (Table 2). Among collagen modifying enzymes, the entire P3H complex is affected with mutations in *CRTAP, LEPRE1* or *PPIB* which result in OI [75, 113, 127, 128] (Figure 8). The three proteins assemble in a 1:1:1 ratio and catalyze the 3-hydroxylation of Pro^{986} in collagen α1(I) and Pro^{707} in α2(I). Mutations in *CRTAP* and *LEPRE1* lead to over modification by LH and P4H enzymes [112]. A similar effect has been observed for collagen C-terminal structural defects indicating for a delay in helix-formation [112]. The chaperoning function of CRTAP might be essential for normal chronological collagen folding but the mechanism is not known. Over-modification by hydroxylation and glycosylation has also been observed for *PPIB*-deficiencies [113]. CyPB (encoded by *PPIB*) is a peptidyl-prolyl cis-trans isomerase and isomerase activity is known to be rate-limiting for collagen folding since the conversion from cis-prolines to trans configuration is necessary for proper triple helix formation [129]. CyPB is ubiquitously expressed and independent of CRTAP and P3H1 (encoded by *LEPRE1*) whereas CRTAP and P3H1 are mutually stabilizing in the P3H complex [130]. That explains the similar phenotypes of *CRTAP*- and *LEPRE1*-deficiencies. Besides the P3H complex, mutations in the LH gene *PLOD2* also lead to skeletal conditions resembling OI. Patients with a homozygous frameshift mutation in the alternative exon of *PLOD2* lack expression of telopeptidyl LH2 which results in impaired collagen cross link formation [131]. Yet again, different mutations for *PLOD2* have been identified which do not give genotype – phenotype correlations. Brothers, having the same *PLOD2* mutations show dissimilar phenotypes diagnosed as mild OI for one brother and more severe and with a higher fracture history as a mild form of Bruck syndrome for the other [131]. Bruck syndrome is denoted by association with OI and congenital joint contractures [115].

Recently, mutations in another collagen modifying enzyme were found to result in an OI phenotype. BMP1, the bone morphogenetic protein 1, cleaves the C-terminal propeptide prior to fibril formation. *BMP1*-deficiency is characterized with a similar phenotype like for dominant collagen mutations in the C-propeptide showing impaired collagen secretion and multiple recurrent fractions except that patients with a *BMP1* mutation have high bone mass [119]. The

interplay of bone production by osteoblasts and resorption by osteoclasts is used for therapeutic strategies (Figure 10). Treatment with bisphosphonates, analogues of pyrophosphates, reduces the life span and function of osteoclasts thereby slowing down the process of bone resorption [125]. Hence, the prolonged osteoblasts still produce mutated collagen but less of the 'OI bone' is resorbed. The arisen increase in bone thickness leads to reduction in fracture rate and bone pain [125]. Other therapies and OI management include calcium/vitamin D supplementation, orthopedic surgery and personal designed exercise programs. Currently, no cure for OI exists.

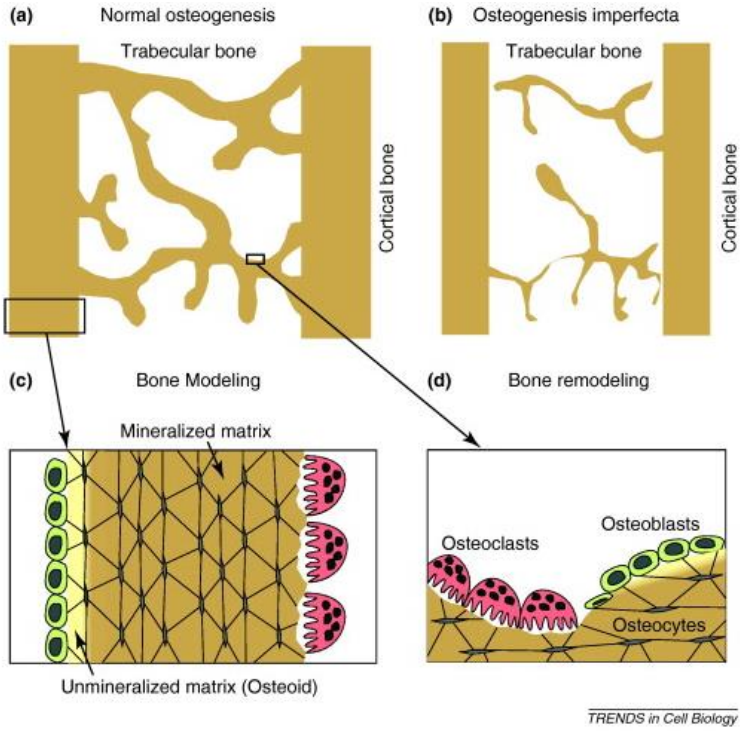


Figure 10| **Bone structure, modeling and remodeling.** (a) Bone consists of a solid outer (cortical) layer and a porous trabecular network filling the cavity inside. (b) In OI, the cortical and trabecular bones are generally thinner. The trabeculae are also less dense and less interconnected. (c) and (d) The lateral growth of cortical bone occurs primarily through modeling: osteoblasts deposit new matrix at the outer surface, which is subsequently mineralized; osteoclasts resorb the inner surface. In OI, imbalance between bone deposition by osteoblasts and resorption by osteoclasts leads to thinner and more porous bones. Taken from [30].

Ehlers-Danlos syndrome

With an incidence of 1:5'000, EDS is amongst the most common heritable connective tissue disorders (Ehlers-Danlos National Foundation, www.ednf.org). EDS is characterized by hyperextensibility of skin and hypermobility of joints. The generalized weakness and fragility of soft connective tissues often goes along with skeletal abnormalities, easy bruising, pronounced bleeding and premature rupture of membranes [132]. Several types of EDS have been classified based on clinical criteria, mode of inheritance and molecular and biochemical analysis (Table 3).

Table 3: The expanded Villefranche classification of Ehlers-Danlos syndrome. Adapted from [132].

EDS type (Villefranche #)	Clinical findings	Inheritance	Gene defects
Classical (I, II)	Skin and joint hypermobility atrophic scars, easy bruising	AD AR	*COL5A1, COL5A2, COL1A1* *TNX-B*
Cardiac-valvular		AR	*COL1A2*
Hypermobility (III)	Joint hypermobility, pain, dislocations	AD	*Unknown, TNX-B*
Vascular (IV)	Thin skin, arterial or uterine rupture, bruising, small joint hyperextensibility	AD	*COL3A1*
Vascular-like		AD	*COL1A1* (R-to-C)
Kyphoscoliosis (VI)	Hypotonia, joint laxity, congenital scoliosis, ocular fragility	AR	*PLOD1*
Musculocontractural		AR	*CHST14*
Progeroid		AR	*B4GALT7, B3GALT6*
Spondylocheirodysplastic		AR	*SLC39A13*
Brittle cornea syndrome		AR	*ZNF469, PRDM5*
Arthrochalasis (VIIa, b)	Severe joint hypermobility, skin mild, scoliosis, bruising	AD	*COL1A1, COL1A2*
Dermatosparaxis (VIIc)	Severe skin fragility, cutis laxa, bruising	AR	*ADAMTS2*

AD = autosomal dominant, AR = autosomal recessive

The classical, hypermobility and vascular types are the most common, whereas the kyphoscoliosis, arthrochalasis and dermatosparaxis types constitute very rare cases [132]. Genetic defects affecting the biosynthesis or enzymatic modification of collagen types I, III and V are the major causes of EDS. Mutations in two of the three genes (*COL5A1* and *COL5A2*) for type V collagen account for about 90% of classical EDS cases [133]. Heterozygous mutations that abolish one *COL5A1* allele result in *COL5A1* haploinsufficiency which yields to half the amount of normal collagen V production. The reduced incorporation of collagen V into collagen I molecules is central in classic EDS pathogenesis. Ultrastructural examination of skin biopsies show irregular loosely packed collagen fibrils which build up the fragile tissue susceptible for atrophic scars.

Until now, homozygous or compound heterozygous mutations of two collagen modifying enzymes, *PLOD1* and *ADAMTS2*, have been described leading to kyphoscoliotic type VI and dermatosparaxis type VIIc EDS, respectively (Table 1 and Table 3). Mutations in *PLOD1* result in weaker collagen structures because Hyl cross-links are more stable than Lys cross-links. *PLOD1*

deficiency accompanied by reduced hydroxylation can be confirmed by analyzing the ratio of lysylpyridinoline (LP) to hydroxylysylpyridinoline (HP) in the urine. A LP/HP ratio of 6 is considered normal, whereas a LP/HP ratio of 1 is low and pathogenic. For a small group of patients with EDS type VI no *PLOD1* mutations could be detected, indicating that mutations for other *PLOD* genes or even the glycosylating enzymes might be the causative agents for the similar phenotype.

Recently, two types of EDS, the musculocontractual and progeroid types have been associated with defects in glycosylating enzymes. Both enzymes are involved in proteoglycan metabolism, one (*CHST14*) is a key sulfotransferase enzyme in the biosynthesis of dermatan sulfate, the other (*B4GALT7*) is involved in the biosynthesis of the glycosaminoglycan (GAG) core. GAG-chain deficiency leads to abnormal collagen fiber assembly [134]. These patients show additional symptoms, like craniofacial abnormalities, joint contractures, wrinkled palms, tapered fingers besides the typical EDS symptoms.

Until now, no defects could be attributed to collagen glycosylation and the collagen glycosylation enzymes. The over modification by hydroxylation and glycosylation seen in OI types VII, VIII and IX and EDS type VI is an effect of the extended time span for the modifying enzyme's activity and the increased substrate availability. Whether mutations in the glycosylating enzymes also contribute to connective tissue disorders is not known. siRNA experiments in the nematode *C.elegans* targeting the *C.elegans* homologue for *GLT25D2* (*D2045.9)* show viable offspring attributed with slow growth, locomotion problems and sterile progeny.

DISCOVERING NEW GLYCOSYLTRANSFERASES *IN SILICO*

In this study we aimed to identify the gene coding for the human collagen glucosyltransferase. The identification of new glycosyltransferases can either be achieved by conventional protein purification methods such as ion exchange chromatography and affinity chromatography from cells, tissues or organisms, or by utilizing bioinformatic tools to limit candidate proteins according to their biochemical properties. This chapter introduces the possibilities of these bioinformatic tools about properties of glycosyltransferases and how they can be used to identify new glycosyltransferases.

The CAZy database and properties of glycosyltransferases

With the human genome sequencing consortium [135] the entire human genome is accessible and all human sequences are available to feed bioinformatic tools for extensive blast searches, comparisons and grouping of proteins. In the field of glycosylation the CAZy-database (Carbohydrate Active Enzymes) provides a growing tool of all carbohydrate interacting proteins across all domains of life [136]. The database describes structurally-related catalytic and

carbohydrate-binding modules of enzymes and groups them into major enzyme classes like glycoside hydrolases, glycosyltransferases, polysaccharide lyases, carbohydrate esterases and auxiliary activities. For glycosyltransferases the grouping in CAZy is structured in families based on amino acid sequence similarities of glycosyltransferases. In 2015, the database describes 97 glycosyltransferase-families with more than 175'000 entries (www.cazy.org). Unlike for glycoside hydrolases which vary a lot in structural properties, most glycosyltransferases accommodate a denoted GT-fold characterized as variation of the Rossmann-like structure (Figure 11). The GT-folds are used to identify and classify the glycosyltransferases into families. Most enzymes with a GT-A fold contain the conserved DxD – motif which is necessary for enzymatic function. One or both of the aspartates are involved in coordinating the metal ion, usually a bivalent cation Mg^{2+} or Mn^{2+}, which contacts with the phosphate group of the bound nucleotide. Unlike the GT-A fold enzymes the GT-B fold enzymes lack the DxD – motif and the nucleotide transfer is metal independent. The mode of interaction with the sugar donor and the GT-B is complex and requires the motion of both Rossmann-like domains of the GT-B fold. For both the GT-A and GT-B fold the three-dimensional structure has been solved (Figure 12). In the last years a third group named GT-C has been introduced which describes the superfamily of integral membrane glycosyltransferases. The GT-C enzymes lack the Rossmann-like fold and contain a modified DxD – motif that can be aligned to ExD, DxE, DDx or DEx.

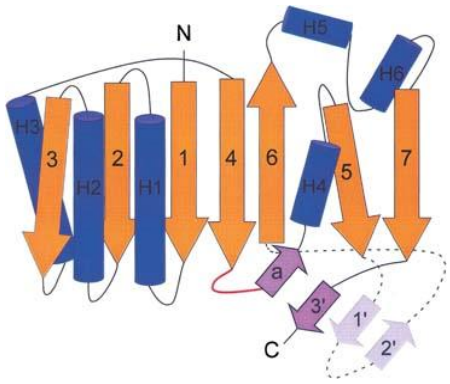


Figure 11| **Topology diagram of the Rossmann-like nucleotide diphospho sugra fold adopted by the GT-A proteins.** Orange arrows indicate β-strands (1–7) forming the main β-sheet of the Rossmann-like fold. Blue cylinders indicate the most conserved α-helices. Magenta arrows indicate β-lip. Red line indicates the location of the S4 loop with DxD motif. Taken from [137].

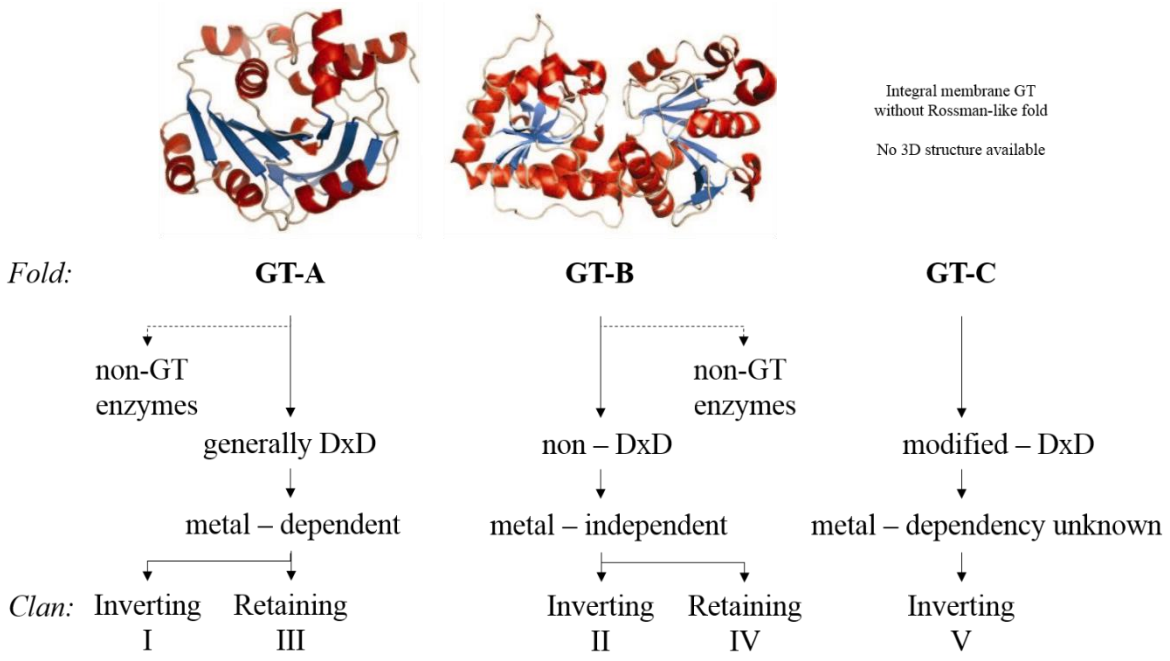


Figure 12| **Structural properties of glycosyltransferases.** Glycosyltransferases accommodate a denoted GT-fold. Accordingly they can be grouped in clans based on the reaction mechanism being either inverting or retaining. Adapted from [137, 138]. Cristal structures are taken from [139].

Another specific feature of GT's is the utilization of an activated donor that can be a nucleoside monophospho sugar, nucleoside diphospho sugar or lipid phospho sugar. The majority of GT's utilizes nucleotide activated monosaccharides as sugar donor. The transfer of the sugar donor to the acceptor site can either be via inversion or retention of the carbohydrate bond (Figure 13). With the expectation that the collagen glucosyltransferase is metal – dependent [46, 140] and transfers the α-*D*-glucose in a retaining mechanism only four GT-families containing human

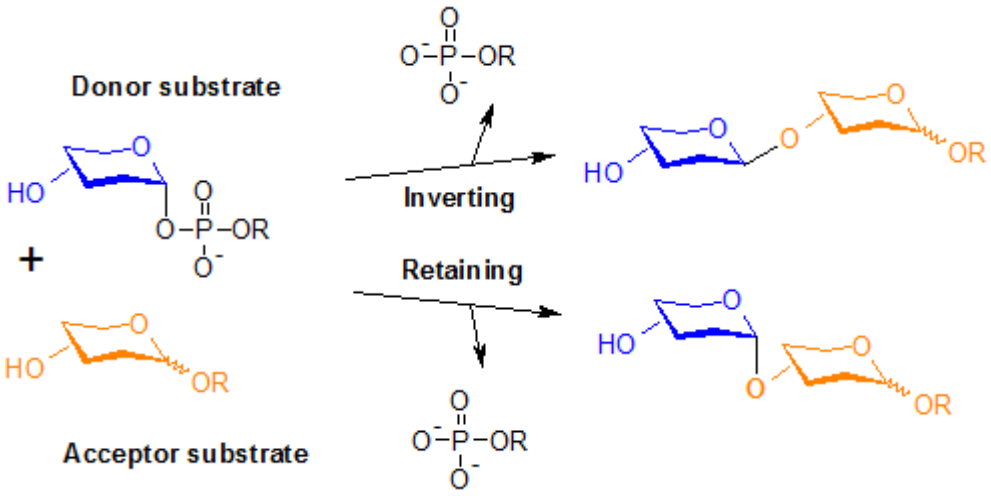


Figure 13| **Mechanism.** Glycosyltransferases transfer the donor glycosyl group with either inversion or retention of the anomeric stereochemistry with respect to the C1 leaving group. Adapted from [139].

sequences would remain to search for possible candidates, namely GT 8, 24, 27, and 64 (Figure 14). The GT 8 family is large and contains 9 human sequences. GT 24 is very distinct and contains only 2. GT 27 and 64 contain 4 and 5 sequences, respectively.

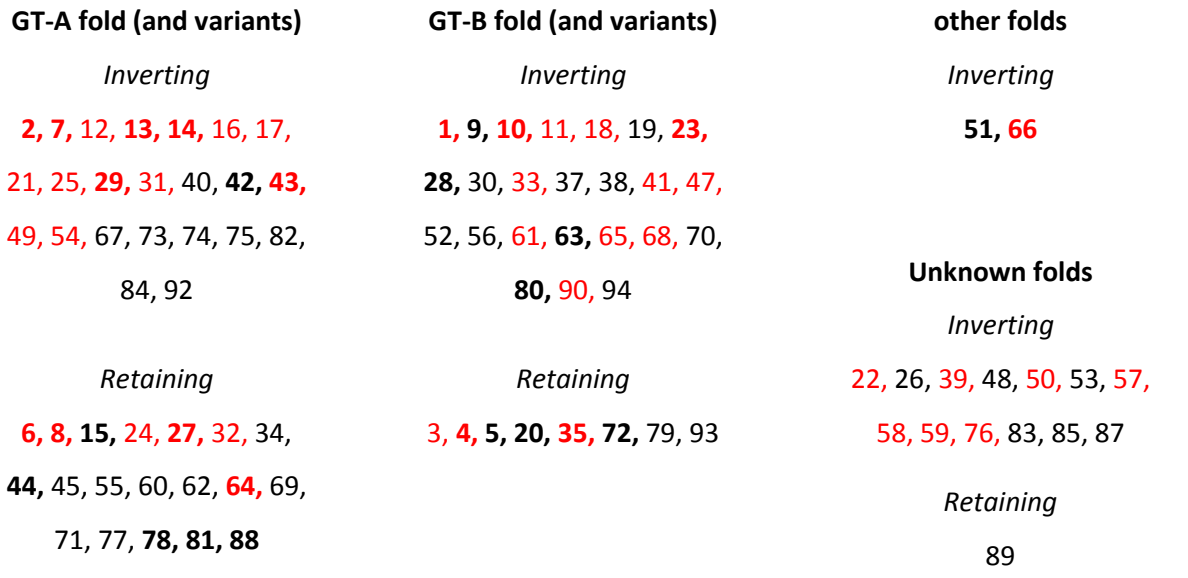


Figure 14| **Classification of CAZy glycosyltransferase families in structural superfamilies.** GT-families in red comprise human sequences, GT-families in bold have at least one member with a solved 3D-structure. Taken from [141].

The knowledge from the *in silico* analysis is also very useful to analyze the output from mass spectrometric reads. Besides the annotated proteins, sequences of unknown or uncharacterized protein hits can be screened for signaling domains to enter the ER and the Golgi apparatus or for the ER-retention signal sequence KDEL since many soluble glycosyltransferases are ER resident as it is expected for the collagen glycosyltransferases [94].

In the end, the combination of the conventional purification protocol with the *in silico* analysis provides a powerful strategy to identify new glycosyltransferases. The identification of the collagen glucosyltransferase would not only unravel the last unknown piece of the human collagen modifying enzyme puzzle but would also allow to characterize the impact of the collagen glycan with respect to its functional role. What is the contribution of the collagen glycan in biosynthesis, assembly, secretion or in the extracellular matrix? Is the glycan important for signaling or collagen remodeling? This is also an important issue for biotechnological applications of industrially-produced collagen. What is the glycan's and enzyme's function and relevance in connective tissue disorder diseases or congenital disorders of glycosylation (CDG)? Would a CDG affect collagen biosynthesis and how?

REFERENCES

1. Bansal, M. and V.S. Ananthanarayanan, *The role of hydroxyproline in collagen folding: conformational energy calculations on oligopeptides containing proline and hydroxyproline.* Biopolymers, 1988. **27**(2): p. 299-312.
2. Bailey, A.J. and C.M. Peach, *Isolation and structural identification of a labile intermolecular crosslink in collagen.* Biochem Biophys Res Commun, 1968. **33**(5): p. 812-9.
3. Bornstein, P. and K.A. Piez, *The nature of the intramolecular cross-links in collagen. The separation and characterization of peptides from the cross-link region of rat skin collagen.* Biochemistry, 1966. **5**(11): p. 3460-73.
4. Katzman, R.L., et al., *Isolation and structure determination of glucosylgalactosylhydroxylysine from sponge and sea anemone collagen.* Biochemistry, 1972. **11**(7): p. 1161-7.
5. Spiro, R.G., *The structure of the disaccharide unit of the renal glomerular basement membrane.* J Biol Chem, 1967. **242**(20): p. 4813-23.
6. Schegg, B., et al., *Core glycosylation of collagen is initiated by two beta(1-O)galactosyltransferases.* Mol Cell Biol, 2009. **29**(4): p. 943-52.
7. Fratzl-Zelman, N., et al., *CRTAP deficiency leads to abnormally high bone matrix mineralization in a murine model and in children with osteogenesis imperfecta type VII.* Bone, 2010. **46**(3): p. 820-6.
8. Cabral, W.A., et al., *Prolyl 3-hydroxylase 1 deficiency causes a recessive metabolic bone disorder resembling lethal/severe osteogenesis imperfecta.* Nat Genet, 2007. **39**(3): p. 359-65.
9. Myllyharju, J. and K.I. Kivirikko, *Collagens, modifying enzymes and their mutations in humans, flies and worms.* Trends Genet, 2004. **20**(1): p. 33-43.
10. Mienaltowski, M.J. and D.E. Birk, *Structure, physiology, and biochemistry of collagens.* Adv Exp Med Biol, 2014. **802**: p. 5-29.
11. Baker, A.T., et al., *Changes in collagen stability and folding in lethal perinatal osteogenesis imperfecta. The effect of alpha 1 (I)-chain glycine-to-arginine substitutions.* Biochem J, 1989. **261**(1): p. 253-7.
12. Cohn, D.H., et al., *Lethal osteogenesis imperfecta resulting from a single nucleotide change in one human pro alpha 1(I) collagen allele.* Proc Natl Acad Sci U S A, 1986. **83**(16): p. 6045-7.
13. Kielty, C.M., I. Hopkinson, and M.E. Grant, *The collagen family: structure, assembly and organization in the extracellular matrix*, in *Connective Tissue and Its Heritable Disorders*, P.M. Royce and B. Steinmann, Editors. 1993, Wiley-Liss. p. 103-147.
14. Myllyharju, J., et al., *Expression of wild-type and modified proalpha chains of human type I procollagen in insect cells leads to the formation of stable [alpha1(I)]2alpha2(I) collagen heterotrimers and [alpha1(I)]3 homotrimers but not [alpha2(I)]3 homotrimers.* J Biol Chem, 1997. **272**(35): p. 21824-30.
15. Boudko, S.P., J. Engel, and H.P. Bachinger, *The crucial role of trimerization domains in collagen folding.* Int J Biochem Cell Biol, 2012. **44**(1): p. 21-32.
16. Koivu, J., *Identification of disulfide bonds in carboxy-terminal propeptides of human type I procollagen.* FEBS Lett, 1987. **212**(2): p. 229-32.
17. Lees, J.F. and N.J. Bulleid, *The role of cysteine residues in the folding and association of the COOH-terminal propeptide of types I and III procollagen.* J Biol Chem, 1994. **269**(39): p. 24354-60.

18. Bachinger, H.P., et al., *Folding mechanism of the triple helix in type-III collagen and type-III pN-collagen. Role of disulfide bridges and peptide bond isomerization.* Eur J Biochem, 1980. **106**(2): p. 619-32.
19. Bachinger, H.P., et al., *Chain assembly intermediate in the biosynthesis of type III procollagen in chick embryo blood vessels.* J Biol Chem, 1981. **256**(24): p. 13193-9.
20. Bateman, J.F., et al., *Characterization of three osteogenesis imperfecta collagen alpha 1(I) glycine to serine mutations demonstrating a position-dependent gradient of phenotypic severity.* Biochem J, 1992. **288 (Pt 1)**: p. 131-5.
21. Chessler, S.D., G.A. Wallis, and P.H. Byers, *Mutations in the carboxyl-terminal propeptide of the pro alpha 1(I) chain of type I collagen result in defective chain association and produce lethal osteogenesis imperfecta.* J Biol Chem, 1993. **268**(24): p. 18218-25.
22. Pace, J.M., et al., *Disruption of one intra-chain disulphide bond in the carboxyl-terminal propeptide of the proalpha1(I) chain of type I procollagen permits slow assembly and secretion of overmodified, but stable procollagen trimers and results in mild osteogenesis imperfecta.* J Med Genet, 2001. **38**(7): p. 443-9.
23. Pace, J.M., et al., *Defective C-propeptides of the proalpha2(I) chain of type I procollagen impede molecular assembly and result in osteogenesis imperfecta.* J Biol Chem, 2008. **283**(23): p. 16061-7.
24. Bonadio, J. and P.H. Byers, *Subtle structural alterations in the chains of type I procollagen produce osteogenesis imperfecta type II.* Nature, 1985. **316**(6026): p. 363-6.
25. Schwarze, U., et al., *Haploinsufficiency for one COL3A1 allele of type III procollagen results in a phenotype similar to the vascular form of Ehlers-Danlos syndrome, Ehlers-Danlos syndrome type IV.* Am J Hum Genet, 2001. **69**(5): p. 989-1001.
26. De Paepe, A., et al., *Mutations in the COL5A1 gene are causal in the Ehlers-Danlos syndromes I and II.* Am J Hum Genet, 1997. **60**(3): p. 547-54.
27. Mitchell, A.L., et al., *Molecular mechanisms of classical Ehlers-Danlos syndrome (EDS).* Hum Mutat, 2009. **30**(6): p. 995-1002.
28. Ishikawa, Y., S. Boudko, and H.P. Bächinger, *Ziploc-ing the structure: Triple helix formation is coordinated by rough endoplasmic reticulum resident PPIases.* Biochim Biophys Acta, 2015.
29. Xu, Y., M. Bhate, and B. Brodsky, *Characterization of the nucleation step and folding of a collagen triple-helix peptide.* Biochemistry, 2002. **41**(25): p. 8143-51.
30. Makareeva, E., N.A. Aviles, and S. Leikin, *Chaperoning osteogenesis: new protein-folding disease paradigms.* Trends Cell Biol, 2011. **21**(3): p. 168-76.
31. Ishida, Y., et al., *Type I collagen in Hsp47-null cells is aggregated in endoplasmic reticulum and deficient in N-propeptide processing and fibrillogenesis.* Mol Biol Cell, 2006. **17**(5): p. 2346-55.
32. Nagata, K., *HSP47 as a collagen-specific molecular chaperone: function and expression in normal mouse development.* Semin Cell Dev Biol, 2003. **14**(5): p. 275-82.
33. Malhotra, V., *COPII vesicles get supersized by ubiquitin.* Cell, 2012. **149**(1): p. 20-1.
34. Saito, K., et al., *TANGO1 facilitates cargo loading at endoplasmic reticulum exit sites.* Cell, 2009. **136**(5): p. 891-902.
35. Jin, L., et al., *Ubiquitin-dependent regulation of COPII coat size and function.* Nature, 2012. **482**(7386): p. 495-500.
36. Kadler, K.E., Y. Hojima, and D.J. Prockop, *Assembly of collagen fibrils de novo by cleavage of the type I pC-collagen with procollagen C-proteinase. Assay of critical*

concentration demonstrates that collagen self-assembly is a classical example of an entropy-driven process. J Biol Chem, 1987. **262**(32): p. 15696-701.
37. Yamauchi, M. and M. Sricholpech, *Lysine post-translational modifications of collagen.* Essays Biochem, 2012. **52**: p. 113-33.
38. Myllyharju, J., *Intracellular Post-Translational Modifications of Collagens*, in *Collagen: Primer in structure, processing and assembly*, J. Brinckmann, H. Notbohm, and P.K. Müller, Editors. 2005, SpringerOnline.
39. Yang, C., et al., *Comprehensive mass spectrometric mapping of the hydroxylated amino acid residues of the α1(V) collagen chain.* J Biol Chem, 2012. **287**(48): p. 40598-610.
40. Weis, M.A., et al., *Location of 3-hydroxyproline residues in collagen types I, II, III, and V/XI implies a role in fibril supramolecular assembly.* J Biol Chem, 2010. **285**(4): p. 2580-90.
41. Dean, D.C., et al., *Isolation of type IV procollagen-like polypeptides from glomerular basement membrane. Characterization of pro-alpha 1(IV).* J Biol Chem, 1983. **258**(1): p. 590-6.
42. Hudson, D.M. and D.R. Eyre, *Collagen prolyl 3-hydroxylation: a major role for a minor post-translational modification?* Connect Tissue Res, 2013. **54**(4-5): p. 245-51.
43. Song, E. and Y. Mechref, *LC-MS/MS identification of the O-glycosylation and hydroxylation of amino acid residues of collagen α-1 (II) chain from bovine cartilage.* J Proteome Res, 2013. **12**(8): p. 3599-609.
44. Inouye, K., et al., *Synthesis and physical properties of (hydroxyproline-proline-glycine)10: hydroxyproline in the X-position decreases the melting temperature of the collagen triple helix.* Arch Biochem Biophys, 1982. **219**(1): p. 198-203.
45. Bann, J.G. and H.P. Bächinger, *Glycosylation/Hydroxylation-induced stabilization of the collagen triple helix. 4-trans-hydroxyproline in the Xaa position can stabilize the triple helix.* J Biol Chem, 2000. **275**(32): p. 24466-9.
46. Kivirikko, K.I. and R. Myllylä, *Collagen glycosyltransferases.* Int Rev Connect Tissue Res, 1979. **8**: p. 23-72.
47. Perdivara, I., M. Yamauchi, and K.B. Tomer, *Molecular Characterization of Collagen Hydroxylysine O-Glycosylation by Mass Spectrometry: Current Status.* Aust J Chem, 2013. **66**(7): p. 760-769.
48. Peterkofsky, B., *Ascorbate requirement for hydroxylation and secretion of procollagen: relationship to inhibition of collagen synthesis in scurvy.* Am J Clin Nutr, 1991. **54**(6 Suppl): p. 1135S-1140S.
49. Friedman, L., et al., *Prolyl 4-hydroxylase is required for viability and morphogenesis in Caenorhabditis elegans.* Proc Natl Acad Sci U S A, 2000. **97**(9): p. 4736-41.
50. Winter, A.D. and A.P. Page, *Prolyl 4-hydroxylase is an essential procollagen-modifying enzyme required for exoskeleton formation and the maintenance of body shape in the nematode Caenorhabditis elegans.* Mol Cell Biol, 2000. **20**(11): p. 4084-93.
51. Holster, T., et al., *Loss of assembly of the main basement membrane collagen, type IV, but not fibril-forming collagens and embryonic death in collagen prolyl 4-hydroxylase I null mice.* J Biol Chem, 2007. **282**(4): p. 2512-9.
52. Berg, R.A. and D.J. Prockop, *The thermal transition of a non-hydroxylated form of collagen. Evidence for a role for hydroxyproline in stabilizing the triple-helix of collagen.* Biochem Biophys Res Commun, 1973. **52**(1): p. 115-20.
53. Holmgren, S.K., et al., *Code for collagen's stability deciphered.* Nature, 1998. **392**(6677): p. 666-7.

54. Engel, J., et al., *The triple helix in equilibrium with coil conversion of collagen-like polytripeptides in aqueous and nonaqueous solvents. Comparison of the thermodynamic parameters and the binding of water to (L-Pro-L-Pro-Gly)n and (L-Pro-L-Hyp-Gly)n.* Biopolymers, 1977. **16**(3): p. 601-22.
55. DeRider, M.L., et al., *Collagen stability: insights from NMR spectroscopic and hybrid density functional computational investigations of the effect of electronegative substituents on prolyl ring conformations.* J Am Chem Soc, 2002. **124**(11): p. 2497-505.
56. Kotch, F.W., I.A. Guzei, and R.T. Raines, *Stabilization of the collagen triple helix by O-methylation of hydroxyproline residues.* J Am Chem Soc, 2008. **130**(10): p. 2952-3.
57. Pokidysheva, E., et al., *Biological role of prolyl 3-hydroxylation in type IV collagen.* Proc Natl Acad Sci U S A, 2014. **111**(1): p. 161-6.
58. Notbohm, H., et al., *Recombinant human type II collagens with low and high levels of hydroxylysine and its glycosylated forms show marked differences in fibrillogenesis in vitro.* J Biol Chem, 1999. **274**(13): p. 8988-92.
59. Eyre, D.R. and M.J. Glimcher, *Analysis of a crosslinked peptide from calf bone collagen: evidence that hydroxylysyl glycoside participates in the crosslink.* Biochem Biophys Res Commun, 1973. **52**(2): p. 663-71.
60. Terajima, M., et al., *Glycosylation and cross-linking in bone type I collagen.* J Biol Chem, 2014. **289**(33): p. 22636-47.
61. Parisuthiman, D., et al., *Biglycan modulates osteoblast differentiation and matrix mineralization.* J Bone Miner Res, 2005. **20**(10): p. 1878-86.
62. Eriksen, H.A., et al., *Differently cross-linked and uncross-linked carboxy-terminal telopeptides of type I collagen in human mineralised bone.* Bone, 2004. **34**(4): p. 720-7.
63. Eyre, D.R., *Collagen: molecular diversity in the body's protein scaffold.* Science, 1980. **207**(4437): p. 1315-22.
64. Jurgensen, H.J., et al., *A novel functional role of collagen glycosylation: interaction with the endocytic collagen receptor uparap/ENDO180.* J Biol Chem, 2011. **286**(37): p. 32736-48.
65. Stawikowski, M.J., et al., *Glycosylation modulates melanoma cell alpha2beta1 and alpha3beta1 integrin interactions with type IV collagen.* J Biol Chem, 2014. **289**(31): p. 21591-604.
66. Pokidysheva, E., et al., *Posttranslational modifications in type I collagen from different tissues extracted from wild type and prolyl 3-hydroxylase 1 null mice.* J Biol Chem, 2013. **288**(34): p. 24742-52.
67. Magiorkinis, E., A. Beloukas, and A. Diamantis, *Scurvy: past, present and future.* Eur J Intern Med, 2011. **22**(2): p. 147-52.
68. Vuori, K., et al., *Characterization of the human prolyl 4-hydroxylase tetramer and its multifunctional protein disulfide-isomerase subunit synthesized in a baculovirus expression system.* Proc Natl Acad Sci U S A, 1992. **89**(16): p. 7467-70.
69. Koivu, J. and R. Myllyla, *Protein disulfide-isomerase retains procollagen prolyl 4-hydroxylase structure in its native conformation.* Biochemistry, 1986. **25**(20): p. 5982-6.
70. Vuori, K., et al., *Site-directed mutagenesis of human protein disulphide isomerase: effect on the assembly, activity and endoplasmic reticulum retention of human prolyl 4-hydroxylase in Spodoptera frugiperda insect cells.* EMBO J, 1992. **11**(11): p. 4213-7.

71. Berg, R.A. and D.J. Prockop, *Purification of (14C) protocollagen and its hydroxylation by prolyl-hydroxylase.* Biochemistry, 1973. **12**(18): p. 3395-401.
72. Cardinale, G.J. and S. Udenfriend, *Prolyl hydroxylase.* Adv Enzymol Relat Areas Mol Biol, 1974. **41**(0): p. 245-300.
73. Vranka, J.A., L.Y. Sakai, and H.P. Bachinger, *Prolyl 3-hydroxylase 1, enzyme characterization and identification of a novel family of enzymes.* J Biol Chem, 2004. **279**(22): p. 23615-21.
74. Wassenhove-McCarthy, D.J. and K.J. McCarthy, *Molecular characterization of a novel basement membrane-associated proteoglycan, leprecan.* J Biol Chem, 1999. **274**(35): p. 25004-17.
75. Morello, R., et al., *CRTAP is required for prolyl 3- hydroxylation and mutations cause recessive osteogenesis imperfecta.* Cell, 2006. **127**(2): p. 291-304.
76. Turpeenniemi-Hujanen, T.M., U. Puistola, and K.I. Kivirikko, *Isolation of lysyl hydroxylase, an enzyme of collagen synthesis, from chick embryos as a homogeneous protein.* Biochem J, 1980. **189**(2): p. 247-53.
77. Kellokumpu, S., et al., *Lysyl hydroxylase, a collagen processing enzyme, exemplifies a novel class of luminally-oriented peripheral membrane proteins in the endoplasmic reticulum.* J Biol Chem, 1994. **269**(48): p. 30524-9.
78. Hautala, T., et al., *Cloning of human lysyl hydroxylase: complete cDNA-derived amino acid sequence and assignment of the gene (PLOD) to chromosome 1p36.3----p36.2.* Genomics, 1992. **13**(1): p. 62-9.
79. Yeowell, H.N. and L.C. Walker, *Tissue specificity of a new splice form of the human lysyl hydroxylase 2 gene.* Matrix Biol, 1999. **18**(2): p. 179-87.
80. Ruotsalainen, H., et al., *Characterization of cDNAs for mouse lysyl hydroxylase 1, 2 and 3, their phylogenetic analysis and tissue-specific expression in the mouse.* Matrix Biol, 1999. **18**(3): p. 325-9.
81. Valtavaara, M., et al., *Cloning and characterization of a novel human lysyl hydroxylase isoform highly expressed in pancreas and muscle.* J Biol Chem, 1997. **272**(11): p. 6831-4.
82. Passoja, K., et al., *Cloning and characterization of a third human lysyl hydroxylase isoform.* Proc Natl Acad Sci U S A, 1998. **95**(18): p. 10482-6.
83. Bank, R.A., et al., *Defective collagen crosslinking in bone, but not in ligament or cartilage, in Bruck syndrome: indications for a bone-specific telopeptide lysyl hydroxylase on chromosome 17.* Proc Natl Acad Sci U S A, 1999. **96**(3): p. 1054-8.
84. Steinmann, B., D.R. Eyre, and P. Shao, *Urinary pyridinoline cross-links in Ehlers-Danlos syndrome type VI.* Am J Hum Genet, 1995. **57**(6): p. 1505-8.
85. Gerriets, J.E., S.L. Curwin, and J.A. Last, *Tendon hypertrophy is associated with increased hydroxylation of nonhelical lysine residues at two specific cross-linking sites in type I collagen.* J Biol Chem, 1993. **268**(34): p. 25553-60.
86. Yeowell, H.N. and L.C. Walker, *Mutations in the lysyl hydroxylase 1 gene that result in enzyme deficiency and the clinical phenotype of Ehlers-Danlos syndrome type VI.* Mol Genet Metab, 2000. **71**(1-2): p. 212-24.
87. van der Slot, A.J., et al., *Identification of PLOD2 as telopeptide lysyl hydroxylase, an important enzyme in fibrosis.* J Biol Chem, 2003. **278**(42): p. 40967-72.
88. Heikkinen, J., et al., *Lysyl hydroxylase 3 is a multifunctional protein possessing collagen glucosyltransferase activity.* J Biol Chem, 2000. **275**(46): p. 36158-63.
89. Sricholpech, M., et al., *Lysyl hydroxylase 3 glucosylates galactosylhydroxylysine residues in type I collagen in osteoblast culture.* J Biol Chem, 2011. **286**(11): p. 8846-56.

90. Myllylä, R., et al., *Expanding the lysyl hydroxylase toolbox: new insights into the localization and activities of lysyl hydroxylase 3 (LH3).* J Cell Physiol, 2007. **212**(2): p. 323-9.
91. Rautavuoma, K., et al., *Characterization of three fragments that constitute the monomers of the human lysyl hydroxylase isoenzymes 1-3. The 30-kDa N-terminal fragment is not required for lysyl hydroxylase activity.* J Biol Chem, 2002. **277**(25): p. 23084-91.
92. Grassmann, W. and H. Schleich, *Uber den Kohlenhydratgehalt des Kollagens II.* Biochem. Z., 1935: p. 277320-328.
93. Butler, W.T. and L.W. Cunningham, *Evidence for the linkage of a disaccharide to hydroxylysine in tropocollagen.* J Biol Chem, 1966. **241**(17): p. 3882-8.
94. Liefhebber, J.M., et al., *The human collagen beta(1-O)galactosyltransferase, GLT25D1, is a soluble endoplasmic reticulum localized protein.* BMC Cell Biol, 2010. **11**: p. 33.
95. Rasmussen, M., A. Eden, and L. Bjorck, *SclA, a novel collagen-like surface protein of Streptococcus pyogenes.* Infect Immun, 2000. **68**(11): p. 6370-7.
96. Whatmore, A.M., *Streptococcus pyogenes sclB encodes a putative hypervariable surface protein with a collagen-like repetitive structure.* Microbiology, 2001. **147**(Pt 2): p. 419-29.
97. Raoult, D., et al., *The 1.2-megabase genome sequence of Mimivirus.* Science, 2004. **306**(5700): p. 1344-50.
98. Shah, N., et al., *Exposure to mimivirus collagen promotes arthritis.* J Virol, 2014. **88**(2): p. 838-45.
99. Legendre, M., et al., *Breaking the 1000-gene barrier for Mimivirus using ultra-deep genome and transcriptome sequencing.* Virol J, 2011. **8**: p. 99.
100. Eriksson, M., et al., *Evidence for 4-hydroxyproline in viral proteins. Characterization of a viral prolyl 4-hydroxylase and its peptide substrates.* J Biol Chem, 1999. **274**(32): p. 22131-4.
101. Luther, K.B., et al., *Mimivirus collagen is modified by bifunctional lysyl hydroxylase and glycosyltransferase enzyme.* J Biol Chem, 2011. **286**(51): p. 43701-9.
102. Wang, I.N., et al., *Evidence for virus-encoded glycosylation specificity.* Proc Natl Acad Sci U S A, 1993. **90**(9): p. 3840-4.
103. Van Etten, J.L., *Unusual life style of giant chlorella viruses.* Annu Rev Genet, 2003. **37**: p. 153-95.
104. Parakkottil Chothi, M., et al., *Identification of an L-rhamnose synthetic pathway in two nucleocytoplasmic large DNA viruses.* J Virol, 2010. **84**(17): p. 8829-38.
105. Piacente, F., et al., *Characterization of a UDP-N-acetylglucosamine biosynthetic pathway encoded by the giant DNA virus Mimivirus.* Glycobiology, 2014. **24**(1): p. 51-61.
106. Piacente, F., et al., *Giant DNA virus mimivirus encodes pathway for biosynthesis of unusual sugar 4-amino-4,6-dideoxy-D-glucose (Viosamine).* J Biol Chem, 2012. **287**(5): p. 3009-18.
107. Stein, H., et al., *Production of bioactive, post-translationally modified, heterotrimeric, human recombinant type-I collagen in transgenic tobacco.* Biomacromolecules, 2009. **10**(9): p. 2640-5.
108. Vuorela, A., et al., *Assembly of human prolyl 4-hydroxylase and type III collagen in the yeast pichia pastoris: formation of a stable enzyme tetramer requires coexpression with collagen and assembly of a stable collagen requires coexpression with prolyl 4-hydroxylase.* EMBO J, 1997. **16**(22): p. 6702-12.

109. Winter, A.D., G. McCormack, and A.P. Page, *Protein disulfide isomerase activity is essential for viability and extracellular matrix formation in the nematode Caenorhabditis elegans.* Dev Biol, 2007. **308**(2): p. 449-61.
110. Willaert, A., et al., *Recessive osteogenesis imperfecta caused by LEPRE1 mutations: clinical documentation and identification of the splice form responsible for prolyl 3-hydroxylation.* J Med Genet, 2009. **46**(4): p. 233-41.
111. Mordechai, S., et al., *High myopia caused by a mutation in LEPREL1, encoding prolyl 3-hydroxylase 2.* Am J Hum Genet, 2011. **89**(3): p. 438-45.
112. Barnes, A.M., et al., *Deficiency of cartilage-associated protein in recessive lethal osteogenesis imperfecta.* N Engl J Med, 2006. **355**(26): p. 2757-64.
113. van Dijk, F.S., et al., *PPIB mutations cause severe osteogenesis imperfecta.* Am J Hum Genet, 2009. **85**(4): p. 521-7.
114. Alanay, Y., et al., *Mutations in the gene encoding the RER protein FKBP65 cause autosomal-recessive osteogenesis imperfecta.* Am J Hum Genet, 2010. **86**(4): p. 551-9.
115. Ha-Vinh, R., et al., *Phenotypic and molecular characterization of Bruck syndrome (osteogenesis imperfecta with contractures of the large joints) caused by a recessive mutation in PLOD2.* Am J Med Genet A, 2004. **131**(2): p. 115-20.
116. LEVENE, C.I. and J. GROSS, *Alterations in state of molecular aggregation of collagen induced in chick embryos by beta-aminopropionitrile (lathyrus factor).* J Exp Med, 1959. **110**: p. 771-90.
117. Colige, A., et al., *Human Ehlers-Danlos syndrome type VII C and bovine dermatosparaxis are caused by mutations in the procollagen I N-proteinase gene.* Am J Hum Genet, 1999. **65**(2): p. 308-17.
118. Martínez-Glez, V., et al., *Identification of a mutation causing deficient BMP1/mTLD proteolytic activity in autosomal recessive osteogenesis imperfecta.* Hum Mutat, 2012. **33**(2): p. 343-50.
119. Asharani, P.V., et al., *Attenuated BMP1 function compromises osteogenesis, leading to bone fragility in humans and zebrafish.* Am J Hum Genet, 2012. **90**(4): p. 661-74.
120. Marini, J.C. and A.R. Blissett, *New genes in bone development: what's new in osteogenesis imperfecta.* J Clin Endocrinol Metab, 2013. **98**(8): p. 3095-103.
121. Forlino, A., et al., *New perspectives on osteogenesis imperfecta.* Nat Rev Endocrinol, 2011. **7**(9): p. 540-57.
122. Stacey, A., et al., *Perinatal lethal osteogenesis imperfecta in transgenic mice bearing an engineered mutant pro-alpha 1(I) collagen gene.* Nature, 1988. **332**(6160): p. 131-6.
123. Valadares, E.R., et al., *What is new in genetics and osteogenesis imperfecta classification?* J Pediatr (Rio J), 2014. **90**(6): p. 536-41.
124. Byers, P.H., *Osteogenesis Imperfecta*, in *Connective Tissue and Its Heritable Disorders*, P.M. Royce and B. Steinmann, Editors. 1993, Wiley-Liss. p. 317 - 350.
125. Harrington, J., E. Sochett, and A. Howard, *Update on the evaluation and treatment of osteogenesis imperfecta.* Pediatr Clin North Am, 2014. **61**(6): p. 1243-57.
126. Marini, J.C., et al., *Consortium for osteogenesis imperfecta mutations in the helical domain of type I collagen: regions rich in lethal mutations align with collagen binding sites for integrins and proteoglycans.* Hum Mutat, 2007. **28**(3): p. 209-21.
127. Barnes, A.M., et al., *Lack of cyclophilin B in osteogenesis imperfecta with normal collagen folding.* N Engl J Med, 2010. **362**(6): p. 521-8.
128. Marini, J.C., W.A. Cabral, and A.M. Barnes, *Null mutations in LEPRE1 and CRTAP cause severe recessive osteogenesis imperfecta.* Cell Tissue Res, 2010. **339**(1): p. 59-70.

129. Steinmann, B., P. Bruckner, and A. Superti-Furga, *Cyclosporin A slows collagen triple-helix formation in vivo: indirect evidence for a physiologic role of peptidyl-prolyl cis-trans-isomerase.* J Biol Chem, 1991. **266**(2): p. 1299-303.
130. Chang, W., et al., *Prolyl 3-hydroxylase 1 and CRTAP are mutually stabilizing in the endoplasmic reticulum collagen prolyl 3-hydroxylation complex.* Hum Mol Genet, 2010. **19**(2): p. 223-34.
131. Puig-Hervás, M.T., et al., *Mutations in PLOD2 cause autosomal-recessive connective tissue disorders within the Bruck syndrome--osteogenesis imperfecta phenotypic spectrum.* Hum Mutat, 2012. **33**(10): p. 1444-9.
132. Malfait, F. and A. De Paepe, *The Ehlers-Danlos syndrome.* Adv Exp Med Biol, 2014. **802**: p. 129-43.
133. Symoens, S., et al., *Comprehensive molecular analysis demonstrates type V collagen mutations in over 90% of patients with classic EDS and allows to refine diagnostic criteria.* Hum Mutat, 2012. **33**(10): p. 1485-93.
134. Miyake, N., T. Kosho, and N. Matsumoto, *Ehlers-Danlos syndrome associated with glycosaminoglycan abnormalities.* Adv Exp Med Biol, 2014. **802**: p. 145-59.
135. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
136. Lombard, V., et al., *The carbohydrate-active enzymes database (CAZy) in 2013.* Nucleic Acids Res, 2014. **42**(Database issue): p. D490-5.
137. Liu, J. and A. Mushegian, *Three monophyletic superfamilies account for the majority of the known glycosyltransferases.* Protein Sci, 2003. **12**(7): p. 1418-31.
138. Coutinho, P.M., et al., *An evolving hierarchical family classification for glycosyltransferases.* J Mol Biol, 2003. **328**(2): p. 307-17.
139. Lairson, L.L., et al., *Glycosyltransferases: structures, functions, and mechanisms.* Annu Rev Biochem, 2008. **77**: p. 521-55.
140. Spiro, R.G. and M.J. Spiro, *Studies on the biosynthesis of the hydroxylysine-linked disaccharide unit of basement membranes and collagens. 3. Tissue and subcellular distribution of glycosyltransferases and the effect of various conditions on the enzyme levels.* J Biol Chem, 1971. **246**(16): p. 4919-25.
141. Breton, C., S. Fournel-Gigleux, and M.M. Palcic, *Recent structures, evolution and mechanisms of glycosyltransferases.* Curr Opin Struct Biol, 2012. **22**(5): p. 540-9.

CONGENITAL DISORDERS OF GLYCOSYLATION (PUBLICATION)

Congenital disorders of glycosylation – a concise chart of glycocalyx dysfunction

Thierry Hennet, Jürg Cabalzar

Institute of Physiology, University of Zurich, CH-8057 Zurich, Switzerland

Trends in Biochemical Sciences, 2015 Mar 31. doi: 10.1016. [Epub ahead of print]

Corresponding author:

Thierry Hennet

Institute of Physiology

University of Zürich

Winterthurerstrasse 190

CH-8057 Zürich

Tel: +41 44 635 5080

Fax: +41 44 635 6814

E-mail: thierry.hennet@uzh.ch

Abstract

Glycosylation is a ubiquitous modification of lipids and proteins. Despite the essential contribution of glycoconjugates to the viability of all living organisms, diseases of glycosylation in humans have only been identified over the last decades. The recent development of next-generation DNA sequencing techniques has accelerated the pace of discovery of novel glycosylation defects. The description of multiple mutations across glycosylation pathways has revealed a tremendous diversity of functional impairments bus also pointed to phenotypic similarities emphasizing the interconnected flow of substrates underlying glycan assembly. The current list of 100 known glycosylation disorders provides an overview on the significance of glycosylation in human development and physiology.

Highlights

- Congenital disorders of glycosylation underline the functional significance of glycosylation in human development and physiology.
- The recent application of next-generation sequencing techniques widely expanded the discovery of glycosylation gene defects.
- The extreme clinical variability associated with glycosylation disorders implies that such diseases are currently underdiagnosed.

Introduction

Glycosylation is by far the most complex form of protein [1, 2] and lipid modification [3, 4] in all domains of life. The tremendous diversity of glycoconjugate structures resulting from intricate biosynthetic pathways is a major factor hampering the assignment of functions to glycans chains. Much has been learnt from the study of disrupted glycosylation genes in model organisms, thereby establishing numerous essential contributions of glycans in regulating cell and organ functions [5]. The study of human diseases of glycosylation brings additional insights by providing a more differentiated view on glycan functions. Indeed, most human mutations are hypomorphic, thus causing partial loss of glycosylation reactions that lead to variable clinical manifestations.

Diseases of glycosylation are also referred to as congenital disorders of glycosylation (CDG). Given the heterogeneity of glycans, the clinical scope of CDG is considerable, ranging from nearly normal phenotypes to severe multi-organ dysfunctions causing infantile lethality. CDG are rare diseases. The prevalence among CDG types is very different from one type to another, but is largely unknown. The difficulty in identifying patients is another reason behind the rarity of CDG. Unspecific symptoms and the lack of simple laboratory tests make the recognition of CDG cases extremely challenging. The identification of CDG has long relied on the detection of under-glycosylated serum transferrin by isoelectric focusing [6]. While easy to perform and requiring only few microliters of blood, this test exclusively reveals alterations of N-glycosylation. Similar blood tests have unfortunately not been established to reliably diagnose defects in other classes of glycosylation. The simplicity of the serum transferrin test also explains why disorders of N-glycosylation account for the majority of known CDG.

Recent developments in genome-wide DNA sequencing technology enable the identification of mutations without previous guessing for candidate genes. As in other fields of biology, next-generation sequencing approaches have increased the pace of discovery for new types of CDG [7]. The barrier of 100 genes defects impairing glycosylation has just been passed. These defects encompass nearly all glycosylation pathways and affect different molecular processes from substrate biosynthesis up to protein trafficking [8, 9]. The recent application of unbiased strategies such as exome and whole-genome sequencing have further revealed CDG-causing mutations in genes previously not associated to glycosylation, thereby expanding our view on these complex pathways.

CDG have originally been classified in two groups. So-called CDG type-I included defects of lipid-linked oligosaccharide assembly up to their transfer to asparagine residues on nascent proteins, whereas CDG type-II included defects of N-glycans trimming and elongation as well as defects in any other class of glycosylation [10]. Because several defects affect multiple glycosylation

pathways, the artificial distinction between CDG type-I and -II has been replaced by a flat nomenclature simply associating implied genes with the suffix CDG [11]. Functionally, defects can also be grouped based on their contribution to glycosylation reactions (**Figure 1**). Accordingly, the present review discusses glycosylation disorders through five functional categories, featuring 1) genes encoding glycosyltransferase enzymes, 2) genes involved in donor substrate biosynthesis, 3) genes mediating the translocation of donor substrates, 4) genes regulating glycosyltransferase localization, and 5) genes affecting the homeostasis of secretory organelles.

Glycosyltransferases

The human genome includes close to 200 glycosyltransferase genes [12]. The majority of these glycosyltransferases are transmembrane proteins anchored in the ER and Golgi membranes [13]. Defects of ER glycosyltransferases involved in the assembly of the oligosaccharide $GlcNAc_2Man_9Glc_3$ (**Figure 2**) lead to defects of N-glycosylation resulting in glycoproteins lacking whole N-glycan chains. Depending on the glycoproteins affected, non-occupancy of N-glycosylation sites can impair protein folding, secretion and stability. At the level of the organism, such defects lead to multiple organ dysfunctions. Neurological symptoms are frequent, featuring psychomotor retardation, ataxia, and hypotonia. Liver and cardiac dysfunctions are also frequently observed as well as endocrine disorders, which mainly affect the sexual maturation of female patients [14].

The functional impairments associated with some glycosyltransferase deficiencies reflect the functional relevance of the involved glycoproteins. For example, O-mannosylation [15] is an essential modification of α-dystroglycan ensuring proper interactions between the dystroglycan complex and proteins of the extracellular matrix [16]. Such interactions are essential for the integrity of muscular fibers, for the migration of neurons in the cortex, and for the retinal architecture [17]. Accordingly, the main manifestations of O-mannosylation disorders are muscular degeneration, brain abnormality, and blindness. Clinically, these disorders belong to the congenital muscular dystrophies and are known as Walker-Warburg syndrome, Muscle-Eye-Brain disease, Fukuyama-type congenital muscular dystrophy, and Limb-girdle muscular dystrophy. The most severe cases are usually associated with mutations in the core mannosyltransferase genes *POMT1* [18] and *POMT2* [19] and in the β1-2 GlcNAc-transferase gene *POMGNT1* [20] (**Figure 2**), but other gene defects also account for severe cases of Walker-Warburg syndrome and Muscle-Eye-Brain disease. To date, defects in 12 genes are known to cause congenital muscular dystrophies, although the functions of some of these genes are still unclear. For example, the *FKTN* and *FKRP* genes encode putative glycosyltransferases involved in O-mannosylation, but their exact substrate specificity and activity remain unknown [21].

Another form of O-linked glycosylation is characterized by the addition of fucose (Fuc) to serine and threonine in the context of the epidermal growth factor (EGF)-like domains and thrombospondin-1 (TSP1) domains. Typical acceptor proteins are members of the Notch family including the ligands Jagged and Delta-like, which are signaling proteins involved in morphogenetic processes [22]. Complete deficiency of core O-fucosyltransferases POFUT1 and POFUT2 has not been described yet, but heterozygous mutations in the *POFUT1* gene have been identified in cases of Dowling-Degos disease, an autosomal dominant pigmentation disorder [23]. Furthermore, mutations in the downstream acting glycosyltransferases, i.e. the β1-3 GlcNAc-transferase LFNG and the β1-3 Glc-transferase B3GALTL (**Figure 2**), have been associated with disorders of vertebral segmentation [24] and to multiple developmental defects known as Peters-Plus syndrome [25], respectively.

In general, defects of core glycosyltransferases are more severe than defects of terminal glycosylation. Nevertheless, the severity of the disease and the scope of organ involvements are also influenced by the functional redundancy inherent to specific glycosyltransferase reactions in the biosynthesis of classes of glycosylation. For example, mucin-type O-glycosylation is initiated by a large family of polypeptide GalNAc-transferases [26]. The partial redundancy in this enzyme family prevents a major loss of this type of glycans in humans, which explains why familial tumoral calcinosis is the only known disease of mucin-type O-glycosylation [27]. Mutations in the polypeptide GalNAc-transferase GALNT3 gene impair the glycosylation of the hormone FGF23, which requires O-GalNAc glycans for its intracellualr trafficking in the Golgi apparatus [28]. Loss of FGF23 secretion leads to hyperphosphatemia and tissue calcification, which are the cardinal symptoms of tumoral calcinosis.

Donor substrates

Despite the hundreds of glycosyltransferases expressed in human cells, only eleven building blocks are used as donor substrates for the assembly of all human glycans. These substrates include nine nucleotide-activated sugars and two dolichol-phosphate (P) linked sugars (**Figure 3A)**. Donor substrates are biosynthesized in the cytosol and in the nucleus for CMP-sialic acid (Sia) through multiple steps including interconversion between monosaccharide isomers. Donor substrates are used across classes of glycosylation, meaning that defects in the formation of individual nucleotide-activated sugars have a broad impact on glycan structures and lead to severe multiorgan disorders. However, clinical severity for a given gene defect widely varies based on the level of residual activity enabled by individual mutations. For example, about 100 mutations have been described for the phosphomannomutase *PMM2* gene [29], which represents by far the most frequent form of CDG. Mutations completely abrogating PMM2 activity lead to embryonic lethality [30] whereas point mutations of minimal impact on the enzymatic activity

will only cause mild intellectual disabilities. PMM2 activity mediates the conversion of mannose (Man)-6-P to Man-1-P, which is an early step in the biosynthesis of GDP-Man (**Figure 3B**). GDP-Man is further converted to dolichol-P-Man by an enzymatic complex encoded by the *DPM1*, *DPM2*, and *DPM3* genes. Dolichol-P-Man is a substrate used in N-glycosylation, O-mannosylation, and for the biosynthesis of the glycosylphosphatidylinositol (GPI) anchors. Accordingly, decreased dolichol-P-Man availability causes of range of diseases sharing features of classical N-glycosylation disorders, but also of congenital muscular dystrophies for O-mannosylation defects.

Sometimes, exome sequencing of untyped CDG cases reveals mutations in genes that were previously associated with diseases unrelated to glycosylation. For example, phosphoglucomutase deficiency resulting from mutations in the *PGM1* gene causes glycogen storage disease XIV, characterized by accumulation of glycogen in muscles because of reduced formation of Glc-6-P from Glc-1-P occurring during breakdown of glycogen [31]. The reverse reaction catalyzed by PGM1, i.e. the formation of Glc-1-P from Glc-6-P is also important for the subsequent formation of UDP-Gal utilized for glycan formation (**Figure 3B**). Indeed, mutations in *PGM1* have been identified as causing CDG with multiple clinical involvements such as growth retardation, cleft palate, muscular and cardiac disorders, and liver dysfunction among other manifestations [32].

The study of glycosylation diseases occasionally points to unexpected findings relative to the biological importance of donor substrate biosynthesis. The *GNE* gene encodes the bifunctional enzyme UDP-GlcNAc 2-epimerase/ManNAc kinase, which catalyzes a rate-limiting step in the biosynthesis of Sia [33]. The disruption of the *Gne* gene in mice is embryonic lethal [34], but decreased GNE activity in human beings is mainly associated with adult-onset, progressive limb-girdle muscle weakness with a remarkable sparing of quadriceps muscles [35]. This rather mild disease suggests that Sia can be efficiently salvaged in humans to bypass any defect of biosynthesis.

Because the biosynthetic pathways of most donor substrates are interconnected, it is tempting to circumvent specific defects by increasing the supply of alternative carbohydrates that can be converted to the missing substrate. Unfortunately, such an approach has only been successful to treat the deficiency of Man-P isomerase (MPI), which catalyzes the interconversion of fructose-6-P and Man-6-P (**Figure 3B**). MPI deficiency is mainly a hepatic-intestinal disease and thus lacks the neurological involvement often found in CDG [36]. The decreased formation of Man-6-P accompanying MPI deficiency can be efficiently compensated by dietary Man supplementation, thereby alleviating disease symptoms [37]. Similarly, dietary supplementation with Gal has recently been shown to normalize serum transferrin glycosylation in patients affected of

phosphoglucomutase 1 (PGM1) deficiency [32], suggesting that Gal supplementation may allievate some of the defects associated with the disease.

Localization of donor substrates

Nucleotide-activated sugars are synthesized in the cytosol and nucleus, but need to be transported to the lumen of the ER and Golgi apparatus for glycosylation reactions. Dedicated antiporters mediate the coupled translocation of nucleotide-activated sugars into the organelles and the return of corresponding nucleotide-monophosphates into the cytosol (**Figure 1**). Most antiporters are specific for a given nucleotide-activated sugar, although multi-specific transporters have also been described. For example, SLC35D1 transports UDP-GlcA and UDP-GalNAc into the ER in exchange of UMP returning to the cytosol. These two nucleotide-activated sugars are involved in the biosynthesis of chondroitin sulfate, a main product of proteoglycans secreted by chondrocytes. Mutations in the *SLC35D1* gene cause a skeletal disease called Schneckenbecken dysplasia, characterized by severe bone abnormalites leading to neonatal lethality [38]. It is likely that other classes of glycosylation are affected by the decreased transport of UDP-GlcA and UDP-GalNAc, but the extent of such alterations has not been addressed yet.

Additional defects of nucleotide-activated sugar transport have been associated with diseases, of which the symptoms reflect the importance of the implied carbohydrate for specific cellular functions. Mutations in the *SLC35C1* gene encoding a Golgi GDP-Fuc transporter impairs terminal fucosylation, which yields epitopes such as ABO and Lewis blood group antigens [39]. Some of these fucosylated epitopes function as ligands for selectins [40] and thereby participate to leukocyte adhesion and extravasation reactions [41]. Accordingly, the shortage of GDP-Fuc in the Golgi caused by defective transport impairs leukocyte trafficking and leads to increased bacterial infections. In addition, affected patients present with short stature, intellectual disability, and mild facial dysmorphism. Hematologic defects and susceptibility to infections could be reverted by oral supplementation with Fuc [42]. By comparison, mutations in the CMP-Sia transporter gene *SLC35A1* were found in a patient with intellectual impairment, seizures, ataxia, thrombocytopenia, renal and cardiac disorders [43]. A general conclusion about the role of sialylation cannot be drawn from these two cases, but the symptoms confirm the importance of Sia for leukocyte and platelet functions.

Dolichol-linked substrates do not use dedicated transporters to reach the ER lumen but proteins have been described that facilitate the translocation of these substrates across membranes. The first of these proteins is called MPDU1 and is required for making dolichol-P-Man and dolichol-P-Glc available to ER mannosyltransferases and glucosyltransferases [44]. These enzymes mediate the elongation of the dolichol-PP-oligosaccharide substrate for N-glycosylation and participate in O-mannosylation and GPI anchor biosynthesis. The mechanism of MPDU1 action is still unknown,

but mutations in the *MPDU1* gene lead to a form of CDG featuring symptoms typical of N-glycosylation disorders, including psychomotor disability, hypotonia, and seizures [45, 46]. Similar symptoms were also observed in patients harboring mutations in the *RFT1* gene [47], which encodes a protein involved in the translocation of the precursor dolichol-PP-linked $GlcNAc_2Man_5$ from the cytosolic into the luminal side of the ER membrane [48] (**Figure 2**). Defective RFT1 activity results in the accumulation of dolichol-PP-$GlcNAc_2Man_5$, which remains unavailable for further extension by luminally-oriented ER mannosyl- and glucosyltransferases.

Localization of glycosyltransferases

A precise localization of glycosyltransferases is also required for proper glycan maturation in the Golgi apparatus. Some glycosyltransferases concentrate in the cisternae of the cis-Golgi, whereas other accumulate in the trans-Golgi. The mechanisms underlying the distribution of glycosyltransferases are not completely understood, but proteins regulating vesicle transports are involved in the process. The Conserved Oligomeric Golgi (COG) complex orchestrates the recycling of medial- and cis-Golgi resident proteins by acting as a tether to connect COPI vesicles with cis-Golgi membranes [49]. COG defects lead to abnormal glycosylation [50] because of missorting of glycosylation enzymes and sugar transporters [51]. Whereas multiple classes of glycosylation are impaired, COG-related disorders are usually identified by detection of underglycosylated serum transferrin just like defects of N-glycosylation.

To date, mutations in seven out of eight COG subunit genes have been described. The most severe diseases are observed for *COG6*, *COG7* and *COG8* mutations, associated with severe neurological impairment, liver dysfunction, and infantile lethality [52-55]. The identification of milder cases of COG6 and COG7 deficiency harboring different mutations [56, 57] however shows that the severity of the disease does not simply relate to the subunit affected but rather to the capability of forming a fully functional COG complex. Besides the severe diseases observed for *COG6* and *COG7* defects, moderate clinical manifestations have been associated with mutations in *COG1* [58, 59], *COG2* [60], *COG4* [61, 62], and *COG5* [63-65].

COG subunits build a complex of two lobes, including COG1 to COG4 in lobe A and COG5 to COG8 in lobe B (**Figure 4**). In general, defects in lobe A lead to milder disease than defects in lobe B. Lobe A appears to be important for overall Golgi architecture, playing a role in Golgi organization and cis-Golgi sorting [66]. Alterations in lobe A lead to accumulation of late glycosylation enzymes in COG complex vesicles, thereby preventing interaction with their substrate. Lobe B rather mediates vesicular sorting of trans-Golgi enzymes through functional interactions with the tethering and fusion machinery of trans-Golgi cisternae [67, 68]. Accordingly, glycosyltransferases from early Golgi cisternae, such as the β1-2 GlcNAc-transferase *MGAT1*, are more affected by a defect in lobe A [66]. By contrast, Gal-transferases and Sia-transferases residing

in trans-Golgi cisternae are more influenced by lobe B alterations [67]. Furthermore, lobe B deficiency mainly results in altered steady state levels of these enzymes due to their translocation to the ER and subsequent proteasomal degradation [65].

Whereas COG defects demonstrate the importance of glycosyltransferase localization for glycosylation, the characterization of another disease called Tn syndrome has pointed to the importance of chaperones in supporting folding and trafficking of specific glycosyltransferases. The Tn syndrome is a clonal defect of core 1 β1-3 Gal-transferase activity limited to a subset of hematopoietic cells. The presentation of bare O-linked GalNAc (the Tn antigen) on erythrocytes leads to the binding of naturally-occurring anti-Tn antibodies and hence to agglutination and hemolysis [69]. Tn antigen presentation on leukocytes and platelets may cause mild leukopenia and thrombocytopenia. Although core 1 β1-3 Gal-transferase activity is decreased in Tn syndrome, no mutations have been found yet in the corresponding *C1GALT1* gene. Rather, mutations in the *COSMC* gene encoding an ER-localized chaperone required for C1GALT1 folding have been identified as causing Tn syndrome [70]. C1GALT1 is the only glycosyltransferase known to undergo chaperone-assisted folding, but the example shows that proper glycosylation also relies on specific proteins such as COSMC that regulate the trafficking of glycosyltransferases from the ER to the Golgi.

Organelle milieu

Glycosyltransferases require co-factors, such as the metal ion Mn^{2+}, and a range of environmental conditions to catalyze glycosylation reactions. The recent application of unbiased genetic approaches, such as homozygosity mapping and exome sequencing, has pointed to novel genes, which affect glycosylation by regulating the acidification and ionic constituents of the secretory pathway.

The *ATP6V0A2* gene encodes a subunit of an H^+-ATPase proton pump localized in the Golgi apparatus [71], which likely regulates pH in Golgi cisternae. Defective ATP6V0A2 action yields structural alterations of Golgi architecture but also causes accumulation of abnormal intracellular vesicles [72]. These changes affect multiple classes of glycosylation as shown by the abnormal N-glycosylation and mucin type O-glycosylation of blood serum proteins. Clinically, mutations in *ATP6V0A2* lead to multiple abnormalities including growth delay and psychomotor disability, but also to skin wrinkling and connective tissue alterations referred to as cutis laxa [73]. Skin and skeletal phenotypes are likely related to alterations of extracellular matrix secretion as indicated by changes of TGF-β signaling observed in affected fibroblasts [71].

Homozygosity mapping and exome sequencing also revealed mutations in the *TMEM165* gene as causing a glycosylation disorder with broad clinical involvement. The five patients identified to

date present with growth and developmental delay, hypotonia, skeletal abnormalities, and hepatomegaly [74]. TMEM165 is a transmembrane protein localized in the Golgi membrane but also found in the plasma membrane, late endosomes and lysosomes. The function of TMEM165 is unclear but appears to be related to transport of calcium [75], which is normally found in high concentrations in the Golgi apparatus [76]. It is unclear whether TMEM165 dysfunctions also affect Mn^{2+} import mediated by the SPCA1 Ca^{2+} pump [77], which would explain the broad glycosylation defect observed in TMEM165 patients.

Exome sequencing will continue to unravel genes of previously unknown or unclear function as cause of CDG. This growing catalog of gene defects will broaden our knowledge on the factors regulating glycosylation, but it will also bring forward new questions regarding the underlying mechanisms of such regulatory pathways. The recent characterization of TMEM165 is a good example outlining the difficult path following the description of mutations in a new gene. This work shows that biochemical and cell biological investigations are always required to understand the biological impact of genetic alterations.

Concluding remarks

This brief review of CDG illustrates the diversity of glycan functions by outlining the widespread consequences of alterations at specific points along biosynthetic pathways. While fascinating, the complexity of CDG and the broad range of disease severity within a CDG type, renders CDG diagnosis challenging. This suggests that CDG is probably underdiagnosed. Accordingly, the application of unbiased sequencing approaches will certainly reveal further gene defects as cause of CDG, but also unravel glycosylation defects in mild disorders such as non-syndromic intellectual disability [78]. Looking back at the evolving CDG landscape of the last decades, it has become clear that the description of these diseases has greatly increased the awareness of the biomedical community for the significance of glycosylation in human development and physiology.

Acknowledgements

We thank Eric Berger for his valuable comments. This work was supported by the Swiss National Foundation grant 310030-149949 to T. Hennet.

Figure legends

Figure 1. Gycosylation reaction. Schematic representation of the key players required for glycosylation reactions occurring in the Golgi apparatus. The biosynthesis of nucleotide-activated sugars (NDP-sugar) takes place in the cytosol whereas glycosyltransferase enzymes are localized on the luminal side of the endomembranes of the secretory pathway. Transporter systems are required for the import of nucleotide-activated sugars into the Golgi apparatus and for maintaining optimal ionic conditions in the organelle, thereby regulating pH, Mn^{2+} import and P export.

Figure 2. Biosynthesis of core structures for N-glycosylation, O-mannosylation, and O-fucosylation. N-glycosylation begins at the ER membrane by the stepwise assembly of dolichol-PP-$GlcNAc_2Man_9Glc_3$, which is transferred to the selected Asn residues of nascent glycoproteins by the oligosaccharyltransferase complex (OST). O-Mannosylated and O-fucosylated glycans are shaped by the sequential addition of different monosaccharides based on the acceptor specificity of glycosyltransferases.

Figure 3. A, List of donor substrates utilized in human cells for glycosylation reactions. B, Biosynthesis pathways of UDP-Gal, UDP-Glc, dolichol-P-Glc (DolP-Glc), GDP-Man, and DolP-Man. The positions of known gene defects are marked with the corresponding gene symbols.

Figure 4. COG organization with display of lobe contributions to organelle architecture and protein trafficking. Both lobes interact with COPI tether proteins (Golgin84, p115) and several SNARE proteins (STX5, STX6, Sly1, GS27, SNAP29). Lobe A and B mediate vesicular retrograde transport of cis (MGAT1, MAN2A1, ST3GAL5) and medial (B4GALT1, ST3GAL1, ST6GAL1) Golgi enzymes, respectively.

Figure 1

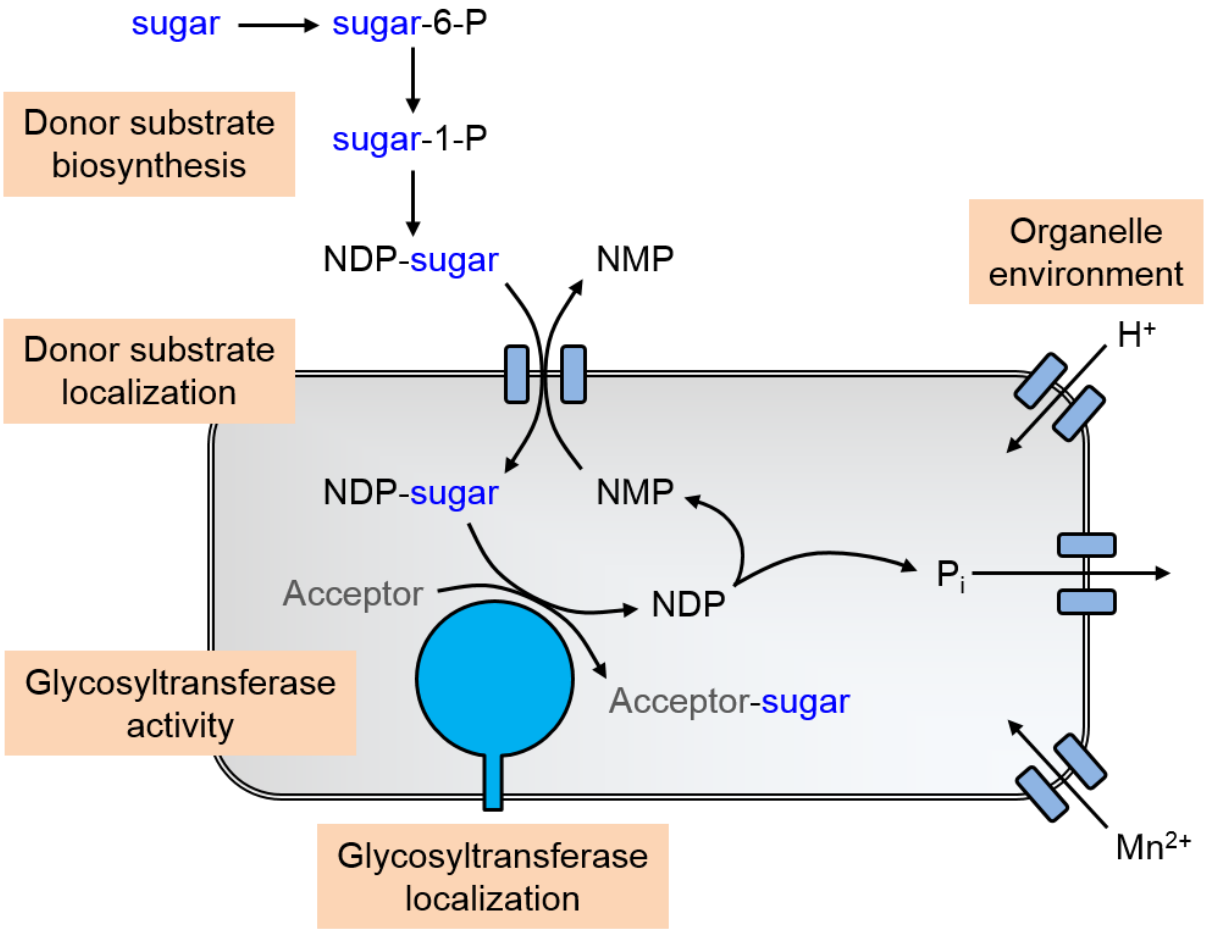

sugar
sugar-6-P
sugar-1-P
NDP-sugar
NMP
Donor substrate biosynthesis
Donor substrate localization
Organelle environment
H+
NDP-sugar
NMP
Acceptor
NDP
Pi
Glycosyltransferase activity
Acceptor-sugar
Mn2+
Glycosyltransferase localization

Figure 2

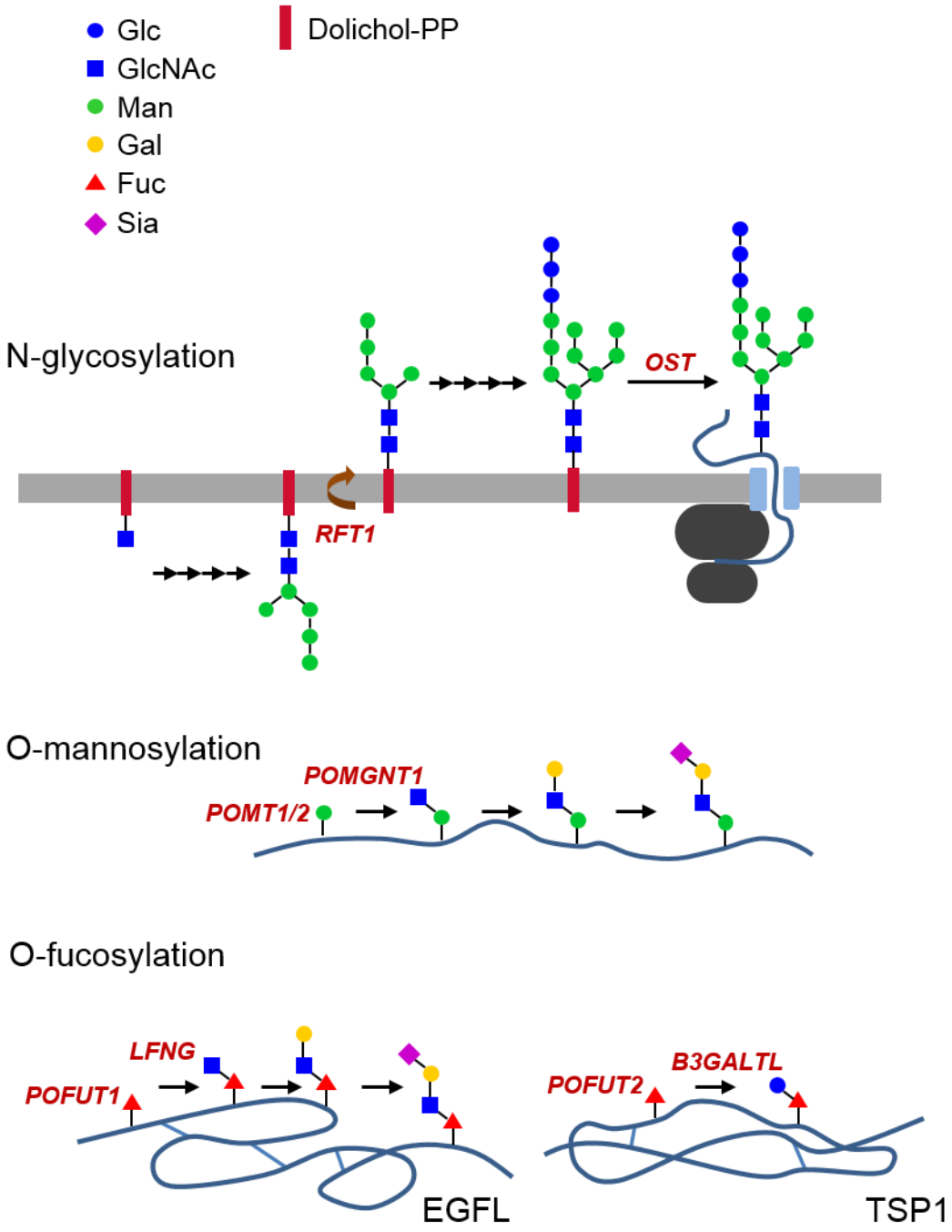

Glc
GlcNAc
Man
Gal
Fuc
Sia
Dolichol-PP
N-glycosylation
OST
RFT1
O-mannosylation
POMGNT1
POMT1/2
O-fucosylation
LFNG
POFUT1
EGFL
B3GALTL
POFUT2
TSP1

Figure 3

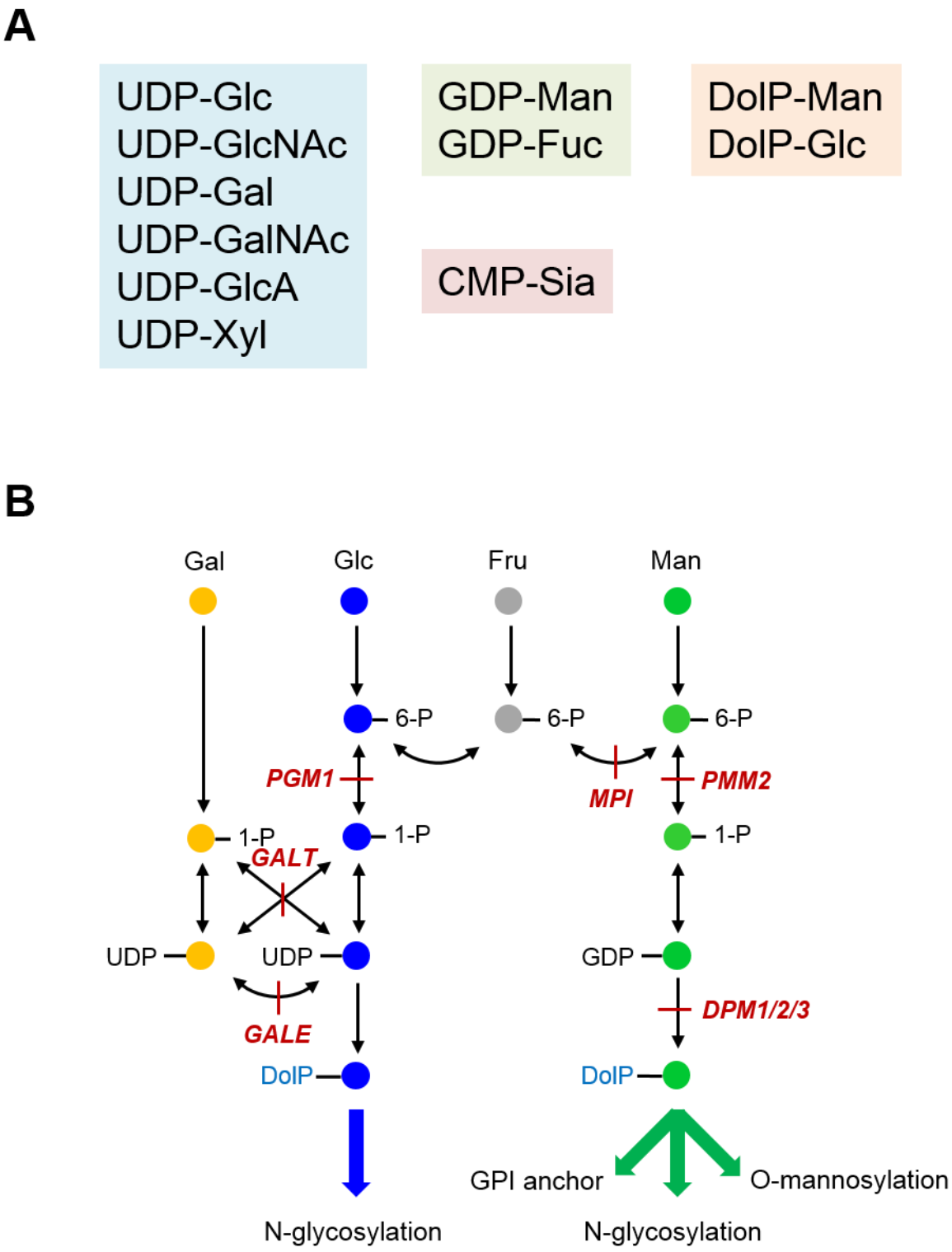
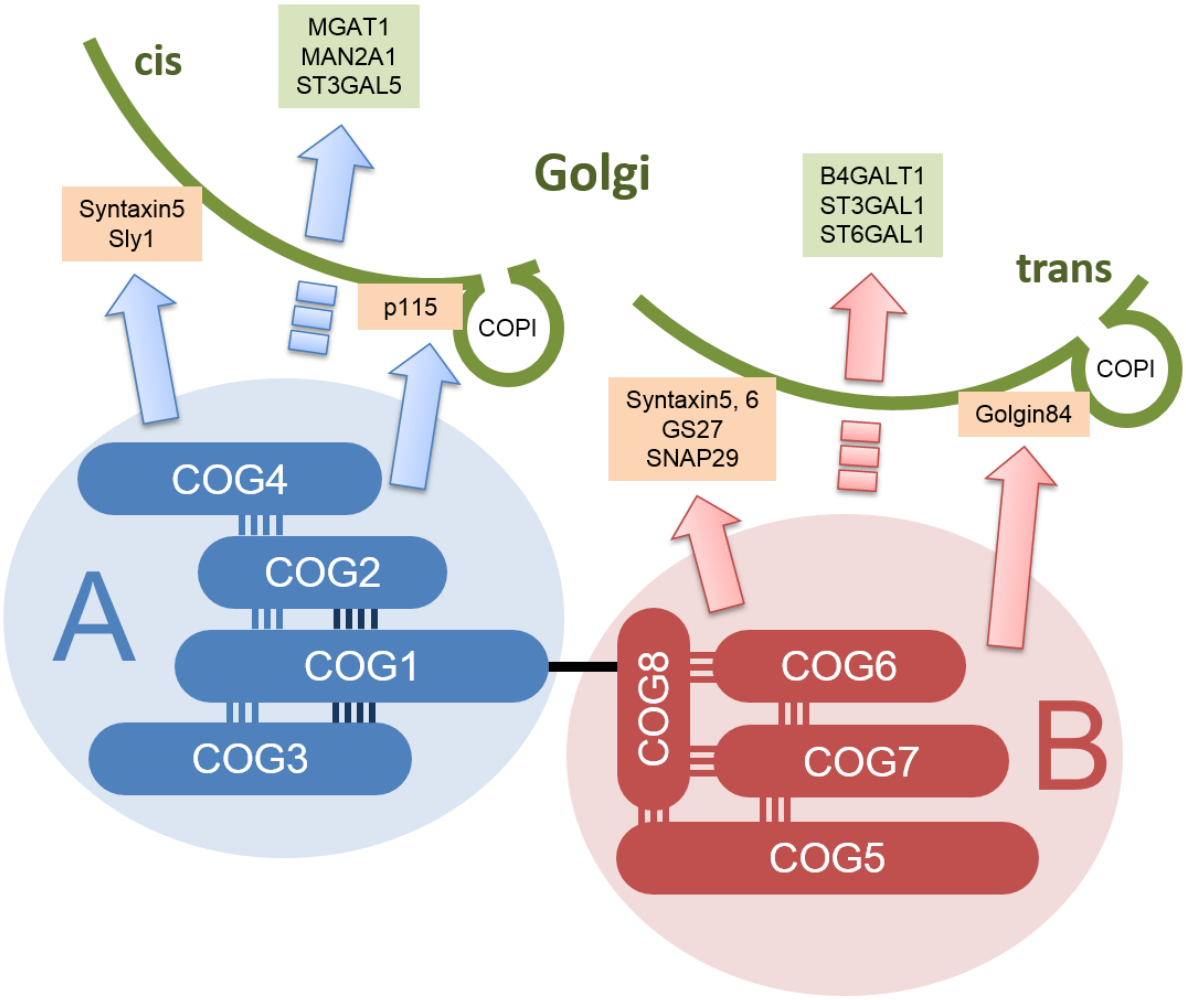

A
UDP-Glc
UDP-GlcNAc
UDP-Gal
UDP-GalNAc
UDP-GlcA
UDP-Xyl
GDP-Man
GDP-Fuc
CMP-Sia
DolP-Man
DolP-Glc
B
Gal
Glc
Fru
Man
6-P
6-P
6-P
PGM1
MPI
PMM2
1-P
1-P
1-P
GALT
UDP
UDP
GDP
GALE
DPM1/2/3
DolP
DolP
GPI anchor
O-mannosylation
N-glycosylation
N-glycosylation

Figure 4


cis
MGAT1
MAN2A1
ST3GAL5
Golgi
B4GALT1
ST3GAL1
ST6GAL1
trans
Syntaxin5
Sly1
p115
COPI
COPI
Syntaxin5, 6
GS27
SNAP29
Golgin84
COG4
COG2
COG1
COG3
A
COG8
COG6
COG7
COG5
B

References

1 Cornfield, A. and Berry, M. (2015) Current aspects of eukaryotic glycosylation. *Trends Biochem Sci* in press

2 Turnbull, J.E. (2015) Complexity and functional diversity of glycosaminoglycans: master cell regulators. *Trends Biochem Sci* in press

3 Schengrund, C.L. (2015) Gangliosides: glycosphingolipids essential for normal neural development and function. *Trends Biochem Sci* in press

4 Ledeen, R.W. and Wu, G. (2015) The multi-tasked life of ganglioside GM1, a true factotum of nature. *Trends Biochem Sci* in press

5 Moremen, K.W. *et al.* (2012) Vertebrate protein glycosylation: diversity, synthesis and function. *Nat Rev Mol Cell Biol* 13, 448-462

6 Babovic-Vuksanovic, D. and O'Brien, J.F. (2007) Laboratory diagnosis of congenital disorders of glycosylation type I by analysis of transferrin glycoforms. *Mol Diagn Ther* 11, 303-311

7 Matthijs, G. *et al.* (2013) Approaches to homozygosity mapping and exome sequencing for the identification of novel types of CDG. *Glycoconj J* 30, 67-76

8 Hennet, T. (2012) Diseases of glycosylation beyond classical congenital disorders of glycosylation. *Biochim Biophys Acta* 1820, 1306-1317

9 Freeze, H.H. *et al.* (2014) Solving Glycosylation Disorders: Fundamental Approaches Reveal Complicated Pathways. *Am J Hum Genet* 94, 161-175

10 Aebi, M. *et al.* (1999) Carbohydrate-deficient glycoprotein syndromes become congenital disorders of glycosylation: an updated nomenclature for CDG. First International Workshop on CDGS. *Glycoconj.J.* 16, 669-671

11 Jaeken, J. *et al.* (2009) CDG nomenclature: time for a change! *Biochim Biophys Acta* 1792, 825-826

12 Varki, A. and Marth, J.D. (1995) Oligosaccharides in vertebrate development. *Seminars in Developmental Biology* 6, 127-138

13 Breton, C. *et al.* (2012) Recent structures, evolution and mechanisms of glycosyltransferases. *Curr Opin Struct Biol* 22, 540-549

14 de Zegher, F. and Jaeken, J. (1995) Endocrinology of the carbohydrate-deficient glycoprotein syndrome type 1 from birth through adolescence. *Pediatr.Res.* 37, 395-401

15 Loibl, M. and Strahl, S. (2013) Protein O-mannosylation: what we have learned from baker's yeast. *Biochim Biophys Acta* 1833, 2438-2446

16 Godfrey, C. *et al.* (2011) Dystroglycanopathies: coming into focus. *Curr Opin Genet Dev* 21, 278-285

17 Abad-Rodriguez, J. and Diez-Revuelta, N. (2015) Axon glycoprotein routing in nerve polarity, function and repair. *Trends Biochem Sci* in press

18 Beltran-Valero de Bernabe, D. *et al.* (2002) Mutations in the O-mannosyltransferase gene POMT1 give rise to the severe neuronal migration disorder Walker-Warburg syndrome. *Am J Hum Genet* 71, 1033-1043

19 van Reeuwijk, J. *et al.* (2005) POMT2 mutations cause alpha-dystroglycan hypoglycosylation and Walker Warburg syndrome. *J Med Genet* 42, 907-912

20 Yoshida, A. *et al.* (2001) Muscular dystrophy and neuronal migration disorder caused by mutations in a glycosyltransferase, POMGnT1. *Dev.Cell* 1, 717-724

21 Praissman, J.L. and Wells, L. (2014) Mammalian O-mannosylation pathway: glycan structures, enzymes, and protein substrates. *Biochemistry* 53, 3066-3078

22 Stanley, P. and Okajima, T. (2010) Roles of glycosylation in Notch signaling. *Curr Top Dev Biol* 92, 131-164

23 Li, M. *et al.* (2013) Mutations in POFUT1, encoding protein O-fucosyltransferase 1, cause generalized Dowling-Degos disease. *Am J Hum Genet* 92, 895-903

24 Sparrow, D.B. *et al.* (2006) Mutation of the lunatic fringe gene in humans causes spondylocostal dysostosis with a severe vertebral phenotype. *Am J Hum Genet* 78, 28-37

25 Lesnik Oberstein, S.A. *et al.* (2006) Peters Plus syndrome is caused by mutations in B3GALTL, a putative glycosyltransferase. *Am J Hum Genet* 79, 562-566

26 Bennett, E.P. *et al.* (2012) Control of mucin-type O-glycosylation: a classification of the polypeptide GalNAc-transferase gene family. *Glycobiology* 22, 736-756

27 Topaz, O. *et al.* (2004) Mutations in GALNT3, encoding a protein involved in O-linked glycosylation, cause familial tumoral calcinosis. *Nat Genet* 36, 579-581

28 Kato, K. *et al.* (2006) Polypeptide GalNAc-transferase T3 and familial tumoral calcinosis. Secretion of fibroblast growth factor 23 requires O-glycosylation. *J Biol Chem* 281, 18370-18377

29 Haeuptle, M.A. and Hennet, T. (2009) Congenital disorders of glycosylation: an update on defects affecting the biosynthesis of dolichol-linked oligosaccharides. *Human Mutation* in press

30 Thiel, C. *et al.* (2006) Targeted disruption of the mouse phosphomannomutase 2 gene causes early embryonic lethality. *Mol Cell Biol* 26, 5615-5620

31 Stojkovic, T. *et al.* (2009) Muscle glycogenosis due to phosphoglucomutase 1 deficiency. *N Engl J Med* 361, 425-427

32 Tegtmeyer, L.C. *et al.* (2014) Multiple phenotypes in phosphoglucomutase 1 deficiency. *N Engl J Med* 370, 533-542

33 Keppler, O.T. *et al.* (1999) UDP-GlcNAc 2-epimerase: a regulator of cell surface sialylation. *Science* 284, 1372-1376

34 Schwarzkopf, M. *et al.* (2002) Sialylation is essential for early development in mice. *Proc Natl Acad Sci U S A* 99, 5267-5270

35 Eisenberg, I. *et al.* (2001) The UDP-N-acetylglucosamine 2-epimerase/N-acetylmannosamine kinase gene is mutated in recessive hereditary inclusion body myopathy. *Nat.Genet.* 29, 83-87

36 Pelletier, V.A. *et al.* (1986) Secretory diarrhea with protein-losing enteropathy, enterocolitis cystica superficialis, intestinal lymphangiectasia, and congenital hepatic fibrosis: a new syndrome. *J Pediatr* 108, 61-65

37 Niehues, R. *et al.* (1998) Carbohydrate-deficient glycoprotein syndrome type Ib. Phosphomannose isomerase deficiency and mannose therapy. *J.Clin.Invest.* 101, 1414-1420

38 Hiraoka, S. *et al.* (2007) Nucleotide-sugar transporter SLC35D1 is critical to chondroitin sulfate synthesis in cartilage and skeletal development in mouse and human. *Nat Med* 13, 1363-1367

39 Lubke, T. *et al.* (2001) Complementation cloning identifies CDG-IIc, a new type of congenital disorders of glycosylation, as a GDP-fucose transporter deficiency. *Nat.Genet.* 28, 73-76

40 Gabius, H.J. (2015) The glycobiology of the CD system: a dictionary for translating marker designations into glycan/lectin structure and function. *Trends Biochem Sci* in press

41 Zarbock, A. *et al.* (2011) Leukocyte ligands for endothelial selectins: specialized glycoconjugates that mediate rolling and signaling under flow. *Blood* 118, 6743-6751

42 Marquardt, T. *et al.* (1999) Correction of leukocyte adhesion deficiency type II with oral fucose. *Blood* 94, 3976-3985

43 Mohamed, M. *et al.* (2013) Intellectual disability and bleeding diathesis due to deficient CMP--sialic acid transport. *Neurology* 81, 681-687

44 Ware, F.E. and Lehrman, M.A. (1996) Expression cloning of a novel suppressor of the Lec15 and Lec35 glycosylation mutations of Chinese hamster ovary cells. *J.Biol.Chem.* 271, 13935-13938

45 Schenk, B. *et al.* (2001) *MPDU1* mutations underlie a novel human congenital disorder of glycosylation (CDG), designated type If. *J.Clin.Invest.* 108, 1687-1695

46 Kranz, C. *et al.* (2001) A mutation in the human *MPDU1* gene causes congenital disorder of glycosylation type If (CDG-If). *J.Clin.Invest* 108, 1613-1619

47 Haeuptle, M.A. *et al.* (2008) Human RFT1 deficiency leads to a disorder of N-linked glycosylation. *Am J Hum Genet* 82, 600-606

48 Helenius, J. *et al.* (2002) Translocation of lipid-linked oligosaccharides across the ER membrane requires Rft1 protein. *Nature* 415, 447-450

49 Ungar, D. *et al.* (2002) Characterization of a mammalian Golgi-localized protein complex, COG, that is required for normal Golgi morphology and function. *J Cell Biol* 157, 405-415

50 Kingsley, D.M. *et al.* (1986) Three types of low density lipoprotein receptor-deficient mutant have pleiotropic defects in the synthesis of N-linked, O-linked, and lipid-linked carbohydrate chains. *J Cell Biol* 102, 1576-1585

51 Oka, T. *et al.* (2004) The COG and COPI complexes interact to control the abundance of GEARs, a subset of Golgi integral membrane proteins. *Mol Biol Cell* 15, 2423-2435

52 Wu, X. *et al.* (2004) Mutation of the COG complex subunit gene COG7 causes a lethal congenital disorder. *Nat Med* 10, 518-523

53 Lubbehusen, J. *et al.* (2010) Fatal outcome due to deficiency of subunit 6 of the conserved oligomeric Golgi complex leading to a new type of congenital disorders of glycosylation. *Hum Mol Genet* 19, 3623-3633

54 Kranz, C. *et al.* (2007) COG8 deficiency causes new congenital disorder of glycosylation type IIh. *Hum Mol Genet* 16, 731-741

55 Foulquier, F. *et al.* (2007) A new inborn error of glycosylation due to a Cog8 deficiency reveals a critical role for the Cog1-Cog8 interaction in COG complex formation. *Hum Mol Genet* 16, 717-730

56 Huybrechts, S. *et al.* (2012) Deficiency of Subunit 6 of the Conserved Oligomeric Golgi Complex (COG6-CDG): Second Patient, Different Phenotype. *JIMD Rep* 4, 103-108

57 Zeevaert, R. *et al.* (2009) A new mutation in COG7 extends the spectrum of COG subunit deficiencies. *Eur J Med Genet* 52, 303-305

58 Foulquier, F. *et al.* (2006) Conserved oligomeric Golgi complex subunit 1 deficiency reveals a previously uncharacterized congenital disorder of glycosylation type II. *Proc Natl Acad Sci U S A* 103, 3764-3769

59 Zeevaert, R. *et al.* (2009) Cerebrocostomandibular-like syndrome and a mutation in the conserved oligomeric Golgi complex, subunit 1. *Hum Mol Genet* 18, 517-524

60 Kodera, H. *et al.* (2014) Mutations in COG2 encoding a subunit of the conserved oligomeric golgi complex cause a congenital disorder of glycosylation. *Clinical genetics*

61 Reynders, E. *et al.* (2009) Golgi function and dysfunction in the first COG4-deficient CDG type II patient. *Hum Mol Genet* 18, 3244-3256

62 Ng, B.G. *et al.* (2011) Identification of the first COG-CDG patient of Indian origin. *Mol Genet Metab* 102, 364-367

63 Paesold-Burda, P. *et al.* (2009) Deficiency in COG5 causes a moderate form of congenital disorders of glycosylation. *Hum Mol Genet* 18, 4350-4356

64 Fung, C.W. *et al.* (2012) COG5-CDG with a Mild Neurohepatic Presentation. *JIMD Rep* 3, 67-70

65 Rymen, D. *et al.* (2012) COG5-CDG: expanding the clinical spectrum. *Orphanet J Rare Dis* 7, 94

66 Peanne, R. *et al.* (2011) Differential effects of lobe A and lobe B of the Conserved Oligomeric Golgi complex on the stability of {beta}1,4-galactosyltransferase 1 and {alpha}2,6-sialyltransferase 1. *Glycobiology* 21, 864-876

67 Pokrovskaya, I.D. *et al.* (2011) Conserved oligomeric Golgi complex specifically regulates the maintenance of Golgi glycosylation machinery. *Glycobiology* 21, 1554-1569

68 Willett, R. *et al.* (2013) COG complexes form spatial landmarks for distinct SNARE complexes. *Nat Commun* 4, 1553

69 Berger, E.G. (1999) Tn-syndrome. *Biochim Biophys Acta* 1455, 255-268

70 Ju, T. and Cummings, R.D. (2005) Protein glycosylation: chaperone mutation in Tn syndrome. *Nature* 437, 1252

71 Fischer, B. *et al.* (2012) Further characterization of ATP6V0A2-related autosomal recessive cutis laxa. *Hum Genet* 131, 1761-1773

72 Hucthagowder, V. *et al.* (2009) Loss-of-function mutations in ATP6V0A2 impair vesicular trafficking, tropoelastin secretion and cell survival. *Hum Mol Genet* 18, 2149-2165

73 Kornak, U. *et al.* (2008) Impaired glycosylation and cutis laxa caused by mutations in the vesicular H+-ATPase subunit ATP6V0A2. *Nat Genet* 40, 32-34

74 Foulquier, F. *et al.* (2012) TMEM165 deficiency causes a congenital disorder of glycosylation. *Am J Hum Genet* 91, 15-26

75 Demaegd, D. *et al.* (2013) Newly characterized Golgi-localized family of proteins is involved in calcium and pH homeostasis in yeast and human cells. *Proc Natl Acad Sci U S A* 110, 6859-6864

76 Pizzo, P. *et al.* (2010) The trans-golgi compartment: A new distinct intracellular Ca store. *Commun Integr Biol* 3, 462-464

77 Vanoevelen, J. *et al.* (2005) The secretory pathway Ca2+/Mn2+-ATPase 2 is a Golgi-localized pump with high affinity for Ca2+ ions. *J Biol Chem* 280, 22800-22808

78 Molinari, F. *et al.* (2008) Oligosaccharyltransferase-subunit mutations in nonsyndromic mental retardation. *Am J Hum Genet* 82, 1150-1157

RESULTS

IDENTIFICATION OF COLGLCT – 3 BASIC APPROACHES

After the identification of the ColGalT in our lab [1], we applied a similar approach to identify the second collagen glycosylating enzyme, the ColGlcT. First we performed a conventional purification protocol to specifically enrich the ColGlcT activity from human fibroblast cells. As an alternative enzyme source we partially purified the ColGlcT from chicken embryo. The enriched ColGlcT active fraction was then subjected to mass spectrometry for protein identification. In a second approach, we targeted the collagen lysyl hydroxylase LH3 in order to copurify the associated ColGlcT. As third procedure we searched the glycosyltransferase containing CAZy database for possible candidate genes, which were then subcloned and expressed with a baculovirus/insect cell expression system and tested for ColGlcT activity.

The presentation of the experimental results obtained in this part is structured in four sections:

1a) Preliminary results obtained from a human enzymatic source

1b) Enrichment for procollagen glucosyltransferase activity reveals the putative glycosyltransferases UGGT2 (Manuscript *Cabalzar et al, 2015)*

2) Affinity purification with PLOD3
3) Candidate search from database – expression and activity

1a) Preliminary results obtained from a human enzymatic source

Human ColGlcT binds to anion exchange DEAE-Sepharose at pH 9.0

In order to pre-fractionate the total cell lysate from human fibroblasts prior to protein chromatography we applied an ER-microsomal isolation protocol. Similar to ColGalT we expect ColGlcT to be an ER resident protein [2]. ER-microsomes were purified by ultracentrifugation of cleared total cell lysates after sonication to open the membrane. The ColGlcT activity was measured in the supernatant indicating that the enzyme could be a soluble ER luminal protein. By use of ion exchange chromatography the proteins were separated based on their charge and buffer system applied. To identify the active fractions specific for ColGlcT activity we used the standard glycosyltransferase assay with [^{14}C]-labeled Glc on heat denatured collagen. Human ColGlcT binds to the anion exchange DEAE-sepharose beads and elutes with increasing salt gradient between 170 to 220 mM NaCl (Figure 1). Anion exchange at pH 9.0 was the only type of ion exchange chromatography which allowed elution of ColGlcT. No binding was observed when applying anion exchange at pH 7.0 and 8.0 as well as cation exchange at pH 6.0 and 6.5 (data not shown).

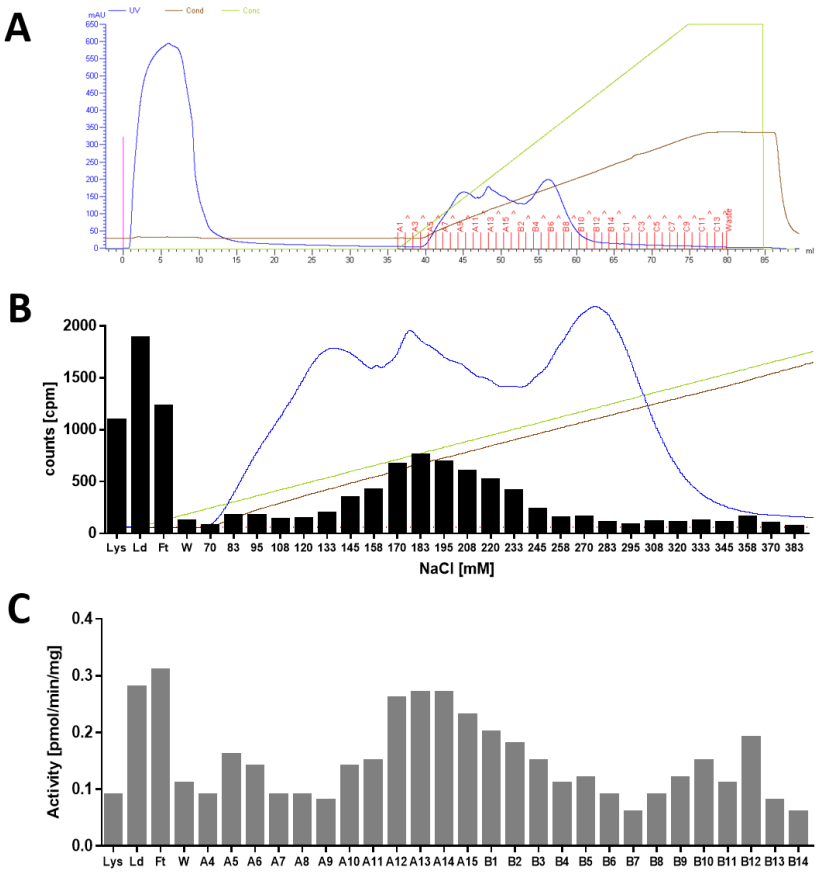


Figure 1| **Elution of human ColGlcT from anion exchange DEAE-sepharose.** Cell lysate from human fibroblasts was ultracentrifuged and run over anion exchange DEAE-sepharose. The proteins elute over a gradient from 20 to 500 mM NaCl (A). The collected fractions were analyzed for collagen glucosyltransferase activity upon transfer of [^{14}C]Glc to collagen (B,C). The highest activity levels for collagen glucosylation were measured in the elution fraction A13 at 180 mM NaCl.

Identification of human collagen modifying proteins by LC-MS/MS

Almost all human collagen modifying enzymes could be detected with mass spectrometric analysis. GLT25D1, LH1-3, P3H, P4Ha, P4Hb and PDIA6 were all identified in the flowthrough fraction after anion exchange DEAE chromatography (Table 1). LH3, P4Hb and PDIA3, 4 and 6 were identified in the eluted fraction with ColGlcT activity. The most prominent glycosyltransferase identified was UDP-glucose:glycoprotein glucosyltransferase 1 (UGGT1). Candidate glycosyltransferases were searched by looking for homology to known glycosyltransferases and for specific structural features like ER-retention signal and DXD hexose binding motifs. None of the putative uncharacterized or unknown proteins could be attributed as a glycosyltransferase. Since the complexity of the eluted sample was very high we aimed for further chromatographic purification experiments. Yet, due to low activity levels, we changed the enzyme source from the human fibroblast cell culture to chicken embryo homogenates.

Table 1. LC-MS/MS protein identification after anion exchange (DEAE) chromatography from human fibroblasts.

Flowthrough DEAE (ingel 1 – 6)			**Elution DEAE (ingel 1 – 6)**		
Description	score[a]	MW[b]	Description	score	MW
Uncharacterized protein C11orf74	35	26	Protein KIAA1199 homolog	32	154
Uncharacterized protein C10orf18 homolog	39	267	UDP-glucose:glycoprotein glucosyltransferase 1	432	178
Neutral alpha-glucosidase	105	107	Neutral alpha-glucosidase	72	107
Leucine-rich repeat-containing protein 47	114	64	WD repeat-containing protein 1	214	67
Procollagen galactosyltransferase 1 (GLT25D1)	50	72	Protein-glutamine gamma-glutamyltransferase 2	116	78
Procollagen-lysine, 2-oxoglutarate 5-dioxygenase 2 (LH2)	115	85	FKPeptidyl-prolyl cis-trans isomerase FKBP9	37	63
Prolyl 3-hydoxylase 1 (P3H1)	97	84	1,4-alpha-glucan-branching enzyme	91	81
Lysyl-tRNA synthetase	51	68	UDP-glucose 4-epimerase	53	39
Procollagen-lysine, 2-oxoglutarate 5-dioxygenase 3 (LH3)	51	85	Procollagen-lysine, 2-oxoglutarate 5-dioxygenase 3 (LH3)	52	85
Procollagen-lysine, 2-oxoglutarate 5-dioxygenase 1 (LH1)	31	84	UTP--glucose-1-phosphate uridylyltransferase	53	57
Transmembrane and TPR repeat-containing protein 3	39	105	Glucose-6-phosphate 1-dehydrogenase	261	59
Protein disulfide-isomerase (P4HB)	38	57	Protein disulfide-isomerase (P4HB)	690	57
Prolyl 4-hydroxylase subunit alpha-1 (P4HA1)	30	61	Inosine-5~-monophosphate dehydrogenase 2	130	56
Protein disulfide-isomerase A6	62	48	Protein disulfide-isomerase A6	935	48
			Protein disulfide-isomerase A4	756	73
			Protein disulfide-isomerase A3	368	57
			UDP-glucose 6-dehydrogenase	148	56
			Putative adenosylhomocysteinase 3	34	58
			UDP-N-acetylhexosamine pyrophosphorylase-like protein 1	84	58
			ERp29	96	29
			Putative uncharacterized protein FLJ32790	42	21

a = Mascot score for protein identification probability. Score > 25 indicates positive identification with p = 0.05
b = molecular weight

1b) Enrichment for procollagen glucosyltransferase activity reveals the putative glycosyltransferase UGGT2 (presented in the manuscript *Cabalzar et al, 2015)*

ENRICHMENT FOR PROCOLLAGEN GLUCOSYLTRANSFERASE ACTIVITY REVEALS THE PUTATIVE GLYCOSYLTRANSFERASE UGGT2

Jürg Cabalzar[1], Andreas J Hülsmeier[1], Peter Gehrig[2], Thierry Hennet[1*]

[1] Institute of Physiology, University of Zurich, Zurich, Switzerland.

[2] Functional Genomics Center Zurich, UZH/ETH Zurich, Zurich, Switzerland.

* Corresponding author

E-mail: thennet@access.uzh.ch

ABSTRACT

Collagen carries a specific type of glycosylation extending on hydroxylysine residues usually by a disaccharide structure hydroxylysine-β1-O-galactose-α1-2-glucose. The core β1-O galactosyltransferases GLT25D1 and GLT25D2 transfer UDP-galactose on hydroxylated lysine residues. The gene encoding the collagen glucosyltransferase enzyme elongating the monosaccharide with α-D-glucose is not known yet. Albeit, the reportedly multifunctional lysyl hydroxylase 3 encoded by plod3 has been found to glucosylate collagen but to a small extent which might not be of sufficient biological relevance and its sole contribution to collagen glucosylation remains questionable. We repeatedly identified the second UDP-glucose:glycoprotein glucosyltransferase homologue UGGT2 in enriched fractions for collagen glucosyltransferase activity. Similar to the collagen glycan structure, UGGT2 is conserved from sponge to human but no function has been assigned for UGGT2 yet.

INTRODUCTION

Collagen, the most abundant protein in our body, encompasses a superfamily of glycoproteins mainly found in the extracellular matrix [1]. Collagen is the major component of skin, bone, cartilage, connective tissues, and basement membranes providing these tissues with high consistency and integrity. The key molecular feature to enable its function is their triple helical form and its aggregation into supramolecular structures. The triple helix formation originates from the amino acid composition of their three α-chains building up the triple helix. The characteristic amino acid repeats Gly-Xaa-Yaa allow a tight helical chain formation with the small amino acid glycine in the center. The Xaa and Yaa positions often represent proline residues, building a polyproline II like helix in each α-chain. The three polyproline II like helices based on the Gly-Xaa-Yaa repeats, are a prerequisite for triple helix formation [2]. Prior to triple helix formation, collagen undergoes a series of post-translational modifications which significantly contribute to collagen's function providing the connective tissue with strength and shape. Among the genomes of vertebrates (and higher invertebrates) are 28 distinct collagen types encoded by at least 45 genes and 23 genes coding for collagen modifying enzymes [1, 3].

The most abundant modification is 4-hydroxylation of proline with about 100 4-hydroxyproline residues (4-Hyp) per 1000 amino acids and about half of the proline residues being hydroxylated in fibrillar type collagens [4]. Unlike 4-hydroxylation, 3-hydroxylation of proline is less abundant with an occurrence of one to two residues per α-chain in collagen types I and II, between three to six residues in types V and XI and more than 10 residues in type IV [5]. The amount of 3-hydroxyproline (3-Hyp) varies not only among different collagen types but also within a certain type, e.g. in type IV collagen the variation goes from one to 20 per 1000 amino acids suggesting

tissue-/cell type-specific differences in modifying enzymes. Mass spectrometric mapping of 3-Hyp in fibrillar type collagens suggests a fundamental role for 3-Hyp in ordered self-assembly of the supramolecular structure, since the 3-Hyp residues were mapped with the characteristic D-periodicity of collagen fibers [6]. In contrast, the 4-Hyp content varies within narrow limit in the same collagen type underlining its critical function for thermal helix stability. Besides the predominant hydroxylation of proline also lysine is hydroxylated forming hydroxylysine residues which can be glycosylated with galactose or galactose-glucose. Galactose (Gal) is linked by the core β-galactosyltransferases *GLT25D1* and *GLT25D2* via an O-glycosidic bond to peptidyl 5-hydroxylysine (Hyl) forming the monosaccharide structure Gal-Hyl [7]. The monosaccharide can be elongated with a glucose (Glc) residue to form the disaccharide structure Glc-Gal-Hyl. The amount and ratio of the disaccharide to the monosaccharide and the free hydroxylysine structure are very variable between different collagen types [8]. The detection of both structures, Gal-Hyl and Glc-Gal-Hyl, at a single lysine residue suggests that glycosylation at this site is dynamic [9, 10]. The mechanism defining the extent of glycosylation is not known. Most of the collagen types carry more glycans of the disaccharide form than the monosaccharide. In mammalian collagen the carbohydrate content of hexoses ranges from 0.4% in skin to about 4% in cartilage and 12% in basement membrane [4]. The tissue dependent difference goes in hand with the predominant type of collagen found in these tissues. Dermal fibrillar type collagen carry fewer glycosylated residues than network forming basement membrane type IV collagen which have a general high degree of post-translational modifications [1, 11].

Alterations of collagen glycosylation have often been reported in several bone and skeletal disorders thereby suggesting that collagen glycosylation might play a role in bone mineralization [12-14]. The glycan might regulate the distribution of bone mineral along the collagen fibril. However, the localization of the glycan to Hyl involved in crosslinking anticipates the functional involvement of the glycan in cross link formation. During cross link maturation the glycan might regulate the cross link species towards either divalent or trivalent cross links depending whether the involved glycan is in the mono- or disaccharide form [15]. Such a mechanism would also explain the variable extent of Hyl glycosylation. Other functions for collagen glycosylation might target the collagen remodeling process. The endocytic collagen receptor μPARAP/Endo180 internalizes collagen for lysosomal degradation via its fibronectin II domain. Additionally, the receptor carries a lectin domain which has been shown to modulate the endocytic efficiency towards highly glycosylated type IV collagen [16]. However, the impact of the glycan is uncertain, since another endocytic receptor, the mannose receptor, does not share this property as the receptor internalizes glycosylated collagen independent of a functional lectin domain [16]. The potential for the glycan to serve as receptor is also described for interactions of cells with basement membranes. The high glycan content in type IV collagen could serve as interaction or

binding receptor in order to recruit cells to the basement membrane. Glycosylated Hyl residues thereby modulate cell adhesion through integrin binding [17]. In contrast, the glycan might have different functions in fibrillar type collagens than the basement membrane type IV collagen. In OI patients, highly glycosylated collagen fibrils show a slight increase in fibril diameter, however it is not clear whether the increased glycosylation or the absence of 3-Hyp due to defective enzymatic activity in these patients is responsible for the disturbance of the lateral fibril growth [18]. In summary, several function for collagen glycosylation including control of matrix mineralization, crosslinking, collagen remodeling, collagen-cell interaction, and fibrillogenesis have been reported. Despite all the reported findings explaining the function of collagen glycosylation, the specific biological function of glycosylated hydroxylysine in relation to the extent and type of glycosylation at their molecular loci, is still not clearly defined.

The importance of the post-translational modifications can be exemplified by the disease scurvy, the old sailor's illness, resulting from a vitamin C (ascorbate) deficiency. People lacking ascorbate show symptoms like easy bruising, fragile capillaries, poor wound healing, skin changes and bone pain. Ascorbate is an important cofactor for the proline and lysine hydroxylase enzymes. Hydroxylation of proline and lysine residues in procollagen α-chains is reduced upon ascorbate deficiency in the cells. The resulting defect in collagen biosynthesis leads to collagen malfunctioning and unstable collagen fibers as observed in scurvy [19]. Many defects in the genes for collagen and its modifying enzymes have been identified leading to Ehlers-Danlos syndrome (EDS), Osteogenesis imperfecta (OI), and Bruck Syndrome (BS) [3, 20-22]. Among the collagen modifying enzymes are defects in the lysyl hydroxylase (LH) genes *PLOD1* and *PLOD2*, the procollagen N- and C-propeptidases *ADAMTS2* and *BMP-1* respectively, and all components of the prolyl 3-hydroxylase complex, namely the peptidyl-prolyl cis-trans isomerase B *PPIB,* the cartilage-associated protein *CRTAP*, and the catalytic Fe- dioxygenase subunit *LEPRE1*. LH-deficiency results in impaired collagen crosslink formation and consequent susceptibility to mechanical disruption of tissue, congenital scoliosis, joint laxity and bone fragility as observed in EDS type VI and BS. Defective ADAMTS2 also lead to autosomal recessive EDS with severe skin fragility, cutis laxa and easy bruising classified in type VIIc. PPIB, CRTAP, and LEPRE1 ensure proper 3-hydroxylation and positional rotation of proline^{986} in collagen type I. Missing 3-Hyp leads to delayed helix formation, following over modification by LH and prolyl 4-hydroxylase resulting in higher mineral content of bone matrix as in OI type VII. Up to now, there is no case reported with a genetic defect in the human prolyl 4-hydroxylase complex leading to the absence of 4-hydroxyproline. Unhydroxylated collagen leads to impaired secretion and its missing deposit in the extracellular space lead to failure of functional connective tissues and basement membrane integrity. RNA-interference studies in C.*elegans* revealed that disruption of the genes encoding for the prolyl 4-hydroxylase complex resulted in embryonic lethality [23-25]. Despite many diseases

have been described for defects in collagen modifying enzymes none has been attributed to a glycosylation defect.

Recently, *GLT25D1* and *GLT25D2*, the genes encoding the collagen galactosyltransferase have been identified [7]. The specific collagen glucosyltransferase (ColGlcT, EC 2.4.1.66) has not been cloned but one of the LH isoenzymes, procollagen-lysine 2-oxoglutarate 5-dioxygenase 3 (PLOD3), has also been shown to possess small amounts of that enzymatic activity [26]. PLOD3 has been postulated to have triple enzymatic activity as in hydroxylation of lysine, galactosylation of hydroxylysine and glucosylation of galactosylhydroxylsine. PLOD3 is a lysine hydroxylase with a Fe- dioxygenase domain at the C-terminus. A possible galactosyltransferase activity residing in the N-terminus could never be confirmed unlike the glucosyltransferase activity, which has been described repeatedly [27, 28]. The contribution of PLOD3 to collagen glucosylation appears to be on a low level compared to ColGalT and it is questionable whether it is of sufficient biological relevance [29]. The collagen glycan is conserved from sponge [30] to human [31] and also in *Gallus gallus* [32] despite the lack of the PLOD3 isoform in *Gallus gallus* indicating the existence of another enzyme for collagen glucosylation.

We purified the ColGlcT activity from chicken embryo, which lack the PLOD3 isoform, by application of a conventional purification protocol. The active fractions were analyzed by tandem mass spectrometry and analyzed for glycosyltransferases. UDP-glucose:glycoprotein glucosyltransferase 2 (UGGT2) was repeatedly identified upon application of different purification protocols.

MATERIALS AND METHODS

Materials

Chicken eggs were purchased from Brüterei Stöckli AG, Ohmstal, Switzerland.

Preparation of gallus ColGlcT protein extract

Protocols for enrichment of collagen glucosyltransferase were adapted from *Myllyla et al.*, *Risteli et al.* and *Schegg et al.*, [7, 33, 34]. 10 day old chicken embryos were homogenized in 225 mM mannitol, 75 mM sucrose, 50 µM dithiothreitol (DTT), and 50 mM Tris-HCl, pH 7.4, at 4 °C and centrifuged at 15,000 × g for 40 min. Supernatants were filtered and proteins were precipitated with 20% $(NH_4)_2SO_4$ at 4 °C (35% saturation). After centrifugation the pellets were dissolved in buffer C1 (200 mM NaCl, 50 µM DTT, 0.05% CHAPS and 50 mM Tris-HCl, pH 7.4, at 4 °C) and dialyzed overnight against two times 5 liters 200 mM NaCl and 50 mM Tris-HOAc, pH 7.4, at 4 °C. The dialyzed fractions were pooled together and centrifuged and the supernatant was used as load for subsequent chromatographic protocols.

Chromatographic steps

ConA-sepharose affinity chromatography

ConA-sepharose beads (GE Healthcare) were equilibrated in buffer C1 supplemented with 1mM $CaCl_2$ and 1 mM $MnCl_2$. The gallus protein extract obtained by $(NH_4)_2SO_4$ precipitation was also supplemented with 1 mM $CaCl_2$ and 1 mM $MnCl_2$ and incubated with the prewashed beads at 4 °C, rotating for 4 h. The bead-protein complex was centrifuged at 1'000 g for 5 min and washed once with supplemented buffer C1. Glycoproteins were eluted with buffer C1 containing 0.5 M α-*D*-methyl-mannopyranoside (Sigma-Aldrich).

Anion exchange chromatography

The protein fraction containing ColGlcT activity eluted from ConA-sepharose chromatography was equilibrated with buffer A1 (50 mM diethanolamine, 20 mM NaCl, pH 9.0). The equilibrated load was applied to prewashed DEAE-sepharose FF 1ml anion exchange column (GE Healthcare) connected to an Äkta FPLC system (GE Healthcare) at a flow rate of 1 ml/min. The loaded column was washed with buffer A1 until the UV-detection level reached baseline again. Proteins were eluted with increasing concentration of buffer A2 (50 mM diethanolamine and 500 mM NaCl, pH 9.0) up to 100% over 40 ml length.

UDP-hexanolamine affinity chromatography

Fractions containing ColGlcT-activity eluted from DEAE- anion exchange column were pooled, diluted with equal volume of buffer U1 (0.15 M NaCl, 10 mM $MnCl_2$, 50 µM DTT, 50 mM Tris-HCl, pH 7.4) and applied on a UDP-hexanolamine-agarose column. The 2 ml bead resin was custom packed in a Benchmark Column 6.6mm/100mm 1xF 1xA (Omnifit) column and connected to an Äkta FPLC system (GE Healthcare). The loaded sample was run at 0.2 ml/min. After washing, the column was eluted with 3 column volumes of buffer U1 containing 5 mM uridine 5′-diphosphate disodium salt hydrate (UDP) (Sigma-Aldrich) and thereafter with 3 column volumes buffer U1 containing 10 mM UDP. During elution 1 ml fractions were collected. The collected fractions were treated with 12.5 units alkaline phosphatase calf intestinal (New England Biolabs) /100 µl eluted fraction for 1 h at 37 °C prior to be used in glycosyltransferase assays.

Gelatin-sepharose affinity chromatography

Fractions containing ColGlcT-activity eluted from ConA-sepharose were pooled and applied on a gelatin-sepharose column. The 5 ml gelatin-sepharose resin (GE Healthcare) was custom packed in a C 16/40 column (GE Healthcare) and connected to an Äkta FPLC system (GE Healthcare). Prior to loading, the column was prewashed with 5 column volumes of buffer G1 (0.2 M NaCl, 0.05% CHAPS and 50 mM Tris-HOAc, pH 7.4 at 4 °C). After loading, the column was washed with high salt buffer G2 (1 M NaCl, 0.05% CHAPS and 50 mM Tris-HOAc, pH 7.4 at 4 °C) and proteins were eluted with low pH buffer G3 (0.2 M NaCl, 0.05% CHAPS and 50 mM Tris-HOAc, pH 5.0 at 4 °C) at a flow

rate of 0.2 ml/min. 1 ml fractions were collected for 20 min. Collected fractions were immediately neutralized with 1 M Tris-HCl pH 8.0.

In a second independent experiment, we additionally added the GT donor substrate UDP to the protocol by supplementing 5 mM UDP (Sigma-Aldrich) to the buffers G1, G2 and G3.

Glycosyltransferase assays

Collagen glycosyltransferase assays were generally performed as described in *Schegg et al.* [7] using [C^{14}] labelled UDP-Glc (20 µCi/ml) (PerkinElmer Life Sciences) and heat denatured collagen type I (250 µg, 10 min at 60 °C) as acceptor. Equal volumes from eluted protein fractions or from total cell lysates of baculoviral-overexpressed proteins (see cloning and expression of ColGlcT candidates) were used as enzyme sources. Assays were incubated at 37° for 3 h and stopped with 5% TCA/ 5% phosphotungstic acid at 4 °C overnight. Precipitated proteins were filtered with Glass Microfibre filters (VWR) and 10 ml IRGA-Safe Plus scintillation fluid (Perkin-Elmer Life Sciences) was added. Radioactivity was measured in a Tri-Carb 2900TR scintillation counter (Perkin-Elmer Life Sciences). Data are shown in counts per minute [cpm] or converted to specific activity [pmol/min/mg total protein] where total protein concentration was determined by Pierce BCA protein assay according to instructor's manual (Thermo Scientific).

UGGT glycosyltransferase assays were performed according to *Trombetta and Parodi* [35]. Similar to the above described assay, also [^{14}C] labelled UDP-Glc (20 µCi/ml) (Perkin-Elmer Life Sciences) and heat denatured acceptor substrates were used. Thyroglobulin, RNaseB and bovine collagen type I (all from Sigma-Aldrich) were heat denatured at 80 °C for 10 min and 200 – 250 µg was used per assay. Additionally, the assays contained 300 nM deoxynojirimycin (Sigma-Aldrich), inhibiting α-glucosidase activity. The assays were incubated at 37 °C for 30 min and stopped with 5% TCA/ 5% phosphotungstic acid at 4 °C overnight. Radioactivity was measured as described above.

Tandem mass spectrometry and data analysis

Collected ColGlcT active fractions were subjected to both protocols for generating tryptic peptides, trypsin in-gel and in-solution digestion protocol. For in-gel digestion, proteins were separated by SDS-PAGE (4 - 12% precast gel, Invitrogen) and the gel lane was divided into 6 regions. The gel pieces from each region were cut in small cubes and destained with 50% methanol in 100 mM ammonium bicarbonate pH 8.0 for 3 h at room temperature. After washing with ammonium bicarbonate, disulfides were reduced with 5 mM tris(carboxyethyl)phosphine hydrochloride (TCEP-HCl, Pierce) in ammonium bicarbonate buffer at 37 °C for 30 min. For alkylation of the cysteine residues, 20 mM iodoacetamide (Sigma-Aldrich) freshly prepared and solubilized in ammonium bicarbonate was added and incubated for 45 min at room temperature

in the dark. After complete removal of iodoacetamide solution the gel pieces were washed once with water and once with ammonium bicarbonate before dehydration in 80% acetonitrile and 20% water. When the gel pieces turned opaque-white, the solution was removed and residual solvent was evaporated in a vacuum concentrator for 5 min. The shrinked gel pieces were covered with 100 ng trypsin in 20 µl Tris buffer (50 mM, pH 8.0) for 10 to 15 min at room temperature. The reswelled gel pieces were fully covered with Tris buffer and digested overnight at 37 °C. For peptide gel extraction, the samples were acidified with trifluoroacetic acid and acetonitrile to a final concentration of 0.1% and 5%, respectively. The samples were sonicated for 5 min in a sonicator water bath. The peptide solution was transferred to new tubes and the gel pieces were incubated with 50% acetonitrile and 0.1% TFA for 20 min at room temperature. After sonication, the peptide solution was combined with the aqueous solution in the new tube and evaporated in a vacuum concentrator until a residual volume of 5 µl. The samples were stored at -20 °C or immediately processed with C18 ZipTip (Millipore) according to the manufacturer's protocol.

For in-solution digestion, eluted fractions were directly processed according to *Shevchenko et al.* [36] following the method from waters (RapiGest SF surfactant protocol). Briefly, after diluting the sample in 100 mM ammonium bicarbonate, 0.1% (w/v) RapiGest (Waters) and 5 mM dithiothreitol (Sigma-Aldrich), the sample was heated for 30 min at 60 °C, cooled, and alkylated in 15 mM iodoacetamide for 30 min in the dark. Proteins were digested with trypsin overnight at 37 °C and acidified with trifluoroacetic acid to a final concentration of 0.5% prior to desalting using a C18 ZipTip (Millipore).

Tryptic digests were subjected to reverse phase LC-MS/MS analysis using a custom packed 150 × 0.075 mm Magic C18- AQ, 3 µm, 200 Å, column (Bischoff GmbH, Leonberg, Germany) and an Orbitrap Velos mass spectrometer (Thermo Scientific). The following 80 min LC gradient was applied: 0 min: 2% buffer B, 50 min: 30% B, 58 min: 50% B, 60 min: 97% B, 70 min: 2% B. Solvent composition of buffer A was 0.2% formic acid and 1% acetonitrile and buffer B contained 0.2% formic acid and 99.8% acetonitrile. Mass spectra were acquired in the m/z range 300 – 1'700 in the Orbitrap mass analyzer at a resolution of 30'000. Spectra were recorded in the collision induced dissociation mode acquiring 10 MS/MS spectra per MS scan with a minimal signal threshold of 2'000 counts. Peptides were identified and assigned using Matrix Science Mascot version 2.4.1.

As an alternative sample preparation method, the eluted protein fractions were alkylated and digested using the Filter aided sample preparation (FASP) protocol from *Wisniewski et al.* [37]. This method is particularly suitable for studying entire proteomes and fractions containing biological membranes.

Cloning and protein expression

The UniprotKB sequences for the ColGlcT candidate genes *UGGT1* and *UGGT2* were purchased from GenScript, New Jersey, NY. The obtained cDNA's were subcloned into pFastBac1 or pFastBac1-Flag (N-terminal Flag-tag) baculovirus transfer vector (Invitrogen). The cDNA of UGGT2 was amplified by PCR with the following primers introducing BamHI and XhoI restriction sites, respectively, sense 5'-TATGGATCCATGGCGCCAGCGAAAGCCACG-3' and antisense 5'-TATCTCGAGTACACCAGTGCTAGAGTTCATCATG-3'. The PCR product was subcloned into pFastBac1 with BamHI/XhoI digestion (all restriction enzymes from New England Biolabs).

The proteins were expressed using the baculovirus expression system (Invitrogen). Briefly, the subcloned pFastBac vectors were transformed into DH10Bac competent *E.coli* for recombinant bacmid DNA production. The recombinant bacmid DNA was transfected with cellfectin into Sf9 cells to generate baculovirus. Baculovirus infected Sf9 cells were harvested after 3 days and lysed in TBS containing 0.5% Triton X-100 (Sigma-Aldrich). The lysates were used for expression verification by Western Blot and ColGlcT activity determination upon baculoviral passage 3.

Western blotting

Total cell lysates from baculoviral infected Sf9 cells were subjected to SDS-PAGE followed by transfer to nitrocellulose membrane (Highbound ECL, GE Healthcare) in semi-dry manner and Western Blot analysis using anti Flag antibody (Sigma-Aldrich). After transfer the membrane was blocked in 3% BSA (Sigma-Aldrich) for 1 h at room temperature. Primary antibody was diluted 1:5'000 in PBS containing 0.5% Tween20 (Sigma-Aldrich) (PBS-T) containing 1% BSA and either incubated at room temperature for 1 h or at 4 °C o/n. The membrane was then washed 3x 10 min in PBS-T followed by incubation with secondary antibody, goat anti rabbit HRP (Sigma-Aldrich), diluted 1:10'000 in PBS-T/ 1% BSA at room temperature for 1 h. The blot was again washed 3x 10 min with PBS-T. Subsequently it was developed with SuperSignal chemiluminescent substrate (Thermo Scientific).

Flag-Tag protein purification

Total cell lysates from baculoviral infected Sf9 cells were incubated with equilibrated ANTI-FLAG M2 affinity gel (Sigma-Aldrich) rotating for 1 h. The purification was performed according to the manufacturer's protocol for the ANTI-FLAG M2 affinity gel. The gel bound FLAG fusion proteins were centrifuged at 1'000 g for 5 min and washed two times with TBS. The protein-bead complex was either directly analyzed by SDS-PAGE and Western blot or the FLAG proteins were eluted with 3X Flag peptide.

Methylation linkage analysis of glucose-galactose-hydroxylysine purified from sponge

Glucose-galactose-hydroxylysine was purified according to *Tenni et al.* [38] and an aliquot was N-acetylated in 70 µl of 0.5 M methanolic HCl:pyridine:acetic anhydride (5:1:1, v:v:v) for 30 min at room temperature. The sample was dried by rotary evaporation and re-evaporated three times from 50 µl methanol. Partially methylated alditol acetates were prepared from the N-acetylated sample and subjected to GC-MS analysis as described before [39].

RESULTS AND DISCUSSION

Isolation of gallus glycoproteins with ConA and anion exchange chromatography

Following the approach for the identification of ColGalT [7, 33], ColGlcT was purified in a similar way from chicken embryo. With the application of a two-step chromatographic protocol, we could specifically enrich the ColGlcT activity by 105-fold (Figure 1). After the initial protein fractionation with ammonium-sulfate, the active fraction was run over ConA-sepharose to isolate glycoproteins and followed by anion exchange chromatography. By screening the eluted fractions for collagen glucosyltransferse activity, the ColGlcT enzyme could be identified to be a glycoprotein and elute at 180 mM NaCl from the anion exchange column at pH 9. The fractions were further separated by SDS-PAGE and divided into gel regions according to molecular weight (Figure 2A). Earlier attempts to purify the ColGlcT suggested a molecular weight of 70 – 80 kDa [40]. In this region we found many ER-chaperones from the heat shock 70-kDa, heat shock cognate 71-kDa and heat shock 90-kDa families (Table 1). By looking at the whole proteome, we found 700 to 900 proteins per fraction and of these 437 proteins were identified in all 3 active fractions (Figure 2B). We identified LH1 and 2 and several glycosyltransferases, among them, UGGT1, UGGT2, glycogen [starch] synthase, glycogen debranching enzyme, beta 1-4 galactosyltransferase 1 and glycogenin. Of these only UGGT1 and LH1 were found to be enriched (Table 2). In general, we still identified 181 proteins that were enriched over the purification process (Figure 2B and Table 2).

No enrichment of gallus ColGlcT with affinity based chromatography

With the application of affinity chromatography, ColGlcT was aimed to specifically enrich upon acceptor or donor interaction. Considering the latter, UDP-hexanolamine serves as donor competitor to UDP-hexoses in the active binding pocket of glycosyltransferases. The UDP-hexanolamine-coupled agarose resin captures UDP-hexose glycosyltransferases but protein identification with mass spectrometry did not reveal any glycosyltransferases in the eluted fractions which goes in hand with the lack of proteins visible by SDS-PAGE (Figure 3B). Most of the proteins run over UDP-hexanolamine agarose did not bind, hence were found in the flowthrough fraction. But ColGlcT activity was measured only from the load fraction but neither from the flowtrough fraction than from the eluates (Figure 3C), suggesting that the ColGlcT still

remained bound. We eluted proteins with an excess of UDP competing for the active binding pocket and therefore not destroying the enzymes activity. However, the ColGlcT activity could not be detected when eluting with 5 mM UDP or with 10 mM UDP. Due to the high amount of UDP in the eluted fraction, we added CIP, an alkaline phosphatase, to eliminate the inhibiting capability of UDP in the ColGlcT enzymatic assay. By adding CIP to the assay, about 50% of the initial activity could be restored (Figure 3D).

Affinity chromatography using gelatin-sepharose as the acceptor resin also failed to isolate the ColGlcT. Unlike the collagen modifying proteins P4H, LEPRE1, CRTAP, and PPIB, the ColGlcT enzyme did not bind to gelatin-sepharose (Figure 3A) [41]. ColGlcT activity was measured in the flowthrough fraction independent whether UDP as donor substrate was present in the chromatography or not (Data not shown).

De novo identification of UGGT, glycogenin and PLOD1 in ColGlcT active fractions

Mass spectrometric analysis of fraction B5 (Figure 1) with highest ColGlcT activity resulted in a similar proteome list (Table 3) like for earlier attempts (Table 1). Additionally, most of the proteins have been identified independent of the alkylation and trypsinizing method applied (Table 3). The glycosyltransferases UGGT and glycogenin are two glucosyltransferases which have been repeatedly identified in ColGlcT enriched fractions (see Tables 1-3). Among glycosyltransferases that have been identified only few times are KDELC1, POGLUT, XXYLT1, β1-4GalT and glycogen debranching enzyme (Gbe). Repeated identification of the UDP-glucose:glycoprotein glucosyltransferase (UGGT) in active fractions after ColGlcT protein purification provoked to have a closer look at this enzyme. UGGT is a well characterized glucosyltransferase which plays a major role in the quality control of N-glycoprotein folding in the ER [42]. Nevertheless, the ability of UGGT to transfer α-linked glucose on the $Man_{7-9}GlcNAc_2$ N-glycan structure of improperly folded proteins makes UGGT a promising candidate to glucosylate collagen as well. The enzyme recognizes hydrophobic patches on acceptor substrates [43, 44], which we also thought could apply for the ColGlcT. Since collagen is glycosylated in the ER before helix formation, the hydrophobic amino acids that would be hidden in the collagen helix center are exposed and could be recognized by the ColGlcT as it is true for UGGT binding hydrophobic patches.

Conserved α1-2 linkage of D-glucose to *O-β*-D-galactopyranosylhydroxylysine from sponge to human

Collagen is highly abundant and widespread in the whole animal kingdom and the glycan composition has always been identified as Glc-Gal-Hyl or Gal-Hyl from sponge to human supposing conservation of the glycan's structural properties. With the identification of UGGT, an α1-3 glucosyltransferase [45], we questioned the conservation of the glucose linkage to be

conserved. In human, the glycosidic linkage has been indirectly determined as α1-2 linkage of D-glucose to O-β-D-galactopyranosylhydroxylysine in renal glomular basement membrane collagen by detecting activity of α-glucosidases towards the disaccharide structure and supported by periodate and methylation studies [31]. To confirm conservation, we purified the Glc-Gal-Hyl from sponge and analyzed the linkage with GC-MS of the methylated hexose fragments. Analysis revealed the expected D-glucose a1-2 D-galactose linkage (Figure 4). Due to the conserved structure, the modifying enzymes would be expected to be so as well. Blast searches revealed homologues proteins for GLT25D2, UGGT1 and UGGT2, and PLOD3 in human, the nematode *C.elegans* and the sponge *A.queenslandica*. However, no PLOD3 was found in *G.gallus*, despite the existence of the Glc-Gal-Hyl glycan.

Recombinant human UGGT1 glucosylate denatured thyroglobulin but not collagen

Two homologues are known for UGGT, UGGT1 and UGGT2. And UGGT1 exists in two isoforms, UGGT1-iso1 and UGGT1-iso2. UGGT1-iso2 lacks 24 amino acids at the N-terminus which is the only difference among the two UGGT1 isoforms. The UGGT proteins are very large with a molecular size of 177 kDa and 1555 amino acid residues. We cloned both isoforms into the baculovirus/insect cell expression system and tested them for ColGlcT activity. Expression of both isoforms was verified with anti-Flag Western blot and could also be visualized by coomassie blue protein staining of SDS-PAGE (Figure 5A and B). Both isoforms exhibit activity towards the denatured form of thyroglobulin but not towards denatured collagen (Figure 5C). Thyroglobulin served as control acceptor protein for UGGT1 activity. UGGT1 glucosylates high mannose glycans of UREA- or heat denatured thyroglobulin but not native thyroglobulin. The activity of UGGT1-lysates towards heat denatured thyroglobulin is 3 times higher than the activity in mock-lysates including endogenous insect UGGT. To increase specificity, we purified UGGT1 from lysates with the Anti-FLAG M2 affinity gel (Figure 5D), but could not measure activity either directly with UGGT1 bound to the beads (Figure 5E) or when UGGT1 was eluted with 3X Flag peptide (data not shown). Both isoforms were also expressed without the Flag-Tag (Figure 5B) and activity on denatured thyroglobulin could be measured but not on denatured collagen (data not shown).

No activity of human UGGT2 on denatured thyroglobulin, RNaseB and collagen

The second homologue UGGT2 shares 55% identity with UGGT1 at the protein level. The highest degree of identity resides in the C-terminal 20% of these proteins, which is denoted as the glycosyltransferase domain. However, only UGGT1 displays the expected functional activity [35, 46]. The function of UGGT2 is unknown and the identification of two homologues for UGGT throughout the entire animal kingdom from sponge to human suggests for highly conserved functions which could be distinct from each other [47]. Indeed, unlike for UGGT1, denatured thyroglobulin is not considered to be a substrate for UGGT2 [46]. Expression of UGGT2 with the

baculoviral/insect cell expression system neither revealed specific activity for denatured collagen nor for denatured thyroglobulin, but we also lacked an expression control (data not shown). To address protein expression, we introduced a Flag-Tag at the N-terminus of the enzyme (Figure 6A). We detected very little to no protein expression at all with anti-Flag Western blot and also with commassie blue staining no band could be visualized when compared to control expression of UGGT1 (Figure 7B). Despite the high similarity of UGGT1 and UGGT2, we could not express full-length UGGT2. Elimination of 220 amino acids at the more variable N-terminus of UGGT2 resulted in strong expression of a 150 kDa UGGT2-short form (Figure 6C). The short form of UGGT2 was not active on denatured collagen and also not on denatured thyroglobulin (Figure 6D). However, the successful expression of UGGT2 after the N-terminal elimination pointed towards a problem at the N-terminus for the full-length UGGT2 expression. Thereof, we designed another construct and eliminated only the signal sequence (S-pep) encoded by the first 25 amino acids (Figure 6A). The resulting construct was named cUGGT2. Both, cUGGT1 and cUGGT2 could be expressed (Figure 7D) and only cUGGT1 exhibited activity towards denatured thyroglobulin and RNaseB (Figure 7A and 7B). But both enzymes did not glucosylate denatured collagen (Figure 7C). Furthermore, cUGGT2 is not active on different pNP-sugar acceptors (Figure 7D).

Purified gallus ColGlcT specifically glucosylates collagen but not thyroglobulin

The fact that UGGT1 and glycogenin were constantly found in active fractions of ColGlcT activity assays does not only make them potential candidates but rather questions the specificity of the collagen-glucosyltransferase assay. To establish clarity we tested the purified gallus ColGlcT on different acceptor proteins. We found that purified gallus ColGlcT specifically glucosylates heat denatured collagen type I in contrast to heat denatured thyroglobulin or water control (Figure 1C). More than three times less counts were measured when no acceptor protein (water control) was added to the assay. The residual activity could be attributed to glycogenin self-glucosylation. To address the contribution of UGGT1 in the assay, we tested the activity of the purified gallus ColGlcT sample on heat denatured thyroglobulin and native thyroglobulin. For both conditions we only measured the residual activity as described before, indicating that the activity on collagen is specific for ColGlcT. This result goes in hand with the lack of baculoviral expressed UGGT1 to glucosylate collagen (Figures 5 and 7).

CONCLUSIONS

Repeated identification of UGGT2 in enriched fractions for ColGlcT activity made UGGT2 a promising candidate as the ColGlcT. Additionally, the existence of two homologues for UGGT throughout the entire animal kingdom from sponge to human indicates highly conserved functions which could be distinct from each other. Supporting evidence derives from siRNA studies in *C.elegans* showing that UGGT2 is essential for viability but not UGGT1 or

Calnexin/Calreticulin [47]. Swapping the carboxy-terminal glycosyltransferase domain from UGGT2 to UGGT1 retained UGGT1 activity but not vice versa. Every other domain exchange towards the N-terminus destroys UGGT1 function [48]. The interchangeable carboxy-terminal glycosyltransferase domain presumes functional glycolytic activity for UGGT2, however the residues defining the acceptor and donor substrate might be located more towards the N-terminal region. Hence, even though UDP-glucose most likely is the donor substrate, we neither identified any glucosyltransferase activity towards different proteinacious acceptor substrates nor towards different *p*NP-sugar substrates. In particular, denatured collagen type I is not a substrate for UGGT2. The published results from a recent study by *Takeda et al.* confirmed our results that UGGT2 is not active on *p*NP-sugar acceptors but they detected very little activity on a proteinacious synthetic IL-8 acceptor and a biosynthetic compound comprised of a BODIPY-dye which was conjugated to $Man_9GlcNAc_2$ via Gly linker [49]. After 6h incubation only 6% of the BODIPY-dye compounds were glucosylated by UGGT2 compared to 38% for UGGT1.

The inability of UGGT2 to glucosylate denatured glycosylated thyroglobulin as well as denatured collagen type I implies that UGGT2 does not accomplish the same function than UGGT1 and UGGT2 does not glucosylate collagen. Whether UGGT2 uses UDP-glucose as donor substrate equally needs to be investigated than what is the specific acceptor substrate. Highly speculative but interesting might be the involvement of UGGT2 as a protein quality control regulator of proteins different than N-glycosylated [50-52].

Among other candidates identified by mass spectrometry, KDELC1 has been cloned and negatively tested for ColGlcT activity (unpublished data), POGLUT1 has been undoubtedly characterized as the protein O-glucosyltransferase RUMI [53], and XXYLT is the elongating xylosyltransferase adding xylose (Xyl) to Xyl-Glc-EGF repeats on the Notch protein [54].

ACKNOWLEDGMENTS

We are grateful to Dr. Simon Barkow at the Functional Genomics Center Zurich for his support with mass spectrometric data analyses, Sacha Schneeberger for the sponge collagen purification and Anna Rommel for revising the manuscript. This work was supported by the University of Zurich and by the Swiss National Foundation grant 310030_149949 to TH.

REFERENCES

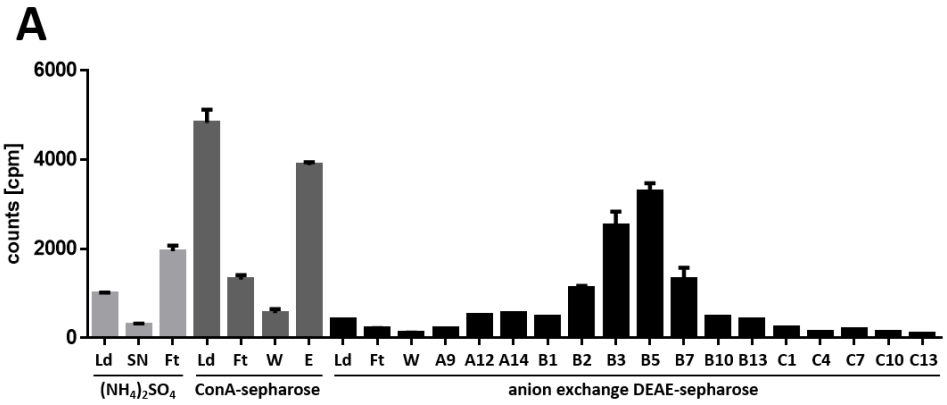
1. Mienaltowski, M.J. and D.E. Birk, *Structure, physiology, and biochemistry of collagens.* Adv Exp Med Biol, 2014. **802**: p. 5-29.
2. Bansal, M. and V.S. Ananthanarayanan, *The role of hydroxyproline in collagen folding: conformational energy calculations on oligopeptides containing proline and hydroxyproline.* Biopolymers, 1988. **27**(2): p. 299-312.

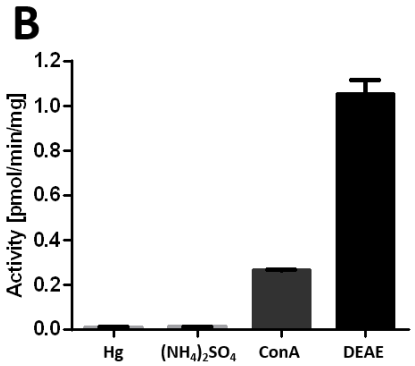
3. Myllyharju, J. and K.I. Kivirikko, *Collagens, modifying enzymes and their mutations in humans, flies and worms.* Trends Genet, 2004. **20**(1): p. 33-43.
4. Kielty, C.M., I. Hopkinson, and M.E. Grant, *The collagen family: structure, assembly and organization in the extracellular matrix*, in *Connective Tissue and Its Heritable Disorders*, P.M. Royce and B. Steinmann, Editors. 1993, Wiley-Liss. p. 103-147.
5. Hudson, D.M. and D.R. Eyre, *Collagen prolyl 3-hydroxylation: a major role for a minor post-translational modification?* Connect Tissue Res, 2013. **54**(4-5): p. 245-51.
6. Weis, M.A., et al., *Location of 3-hydroxyproline residues in collagen types I, II, III, and V/XI implies a role in fibril supramolecular assembly.* J Biol Chem, 2010. **285**(4): p. 2580-90.
7. Schegg, B., et al., *Core glycosylation of collagen is initiated by two beta(1-O)galactosyltransferases.* Mol Cell Biol, 2009. **29**(4): p. 943-52.
8. Kivirikko, K.I. and R. Myllylä, *Collagen glycosyltransferases.* Int Rev Connect Tissue Res, 1979. **8**: p. 23-72.
9. Perdivara, I., M. Yamauchi, and K.B. Tomer, *Molecular Characterization of Collagen Hydroxylysine O-Glycosylation by Mass Spectrometry: Current Status.* Aust J Chem, 2013. **66**(7): p. 760-769.
10. Yang, C., et al., *Comprehensive mass spectrometric mapping of the hydroxylated amino acid residues of the α1(V) collagen chain.* J Biol Chem, 2012. **287**(48): p. 40598-610.
11. Myllyharju, J., *Intracellular Post-Translational Modifications of Collagens*, in *Collagen: Primer in structure, processing and assembly*, J. Brinckmann, H. Notbohm, and P.K. Müller, Editors. 2005, SpringerOnline.
12. Parisuthiman, D., et al., *Biglycan modulates osteoblast differentiation and matrix mineralization.* J Bone Miner Res, 2005. **20**(10): p. 1878-86.
13. Eriksen, H.A., et al., *Differently cross-linked and uncross-linked carboxy-terminal telopeptides of type I collagen in human mineralised bone.* Bone, 2004. **34**(4): p. 720-7.
14. Eyre, D.R., *Collagen: molecular diversity in the body's protein scaffold.* Science, 1980. **207**(4437): p. 1315-22.
15. Terajima, M., et al., *Glycosylation and cross-linking in bone type I collagen.* J Biol Chem, 2014. **289**(33): p. 22636-47.
16. Jurgensen, H.J., et al., *A novel functional role of collagen glycosylation: interaction with the endocytic collagen receptor uparap/ENDO180.* J Biol Chem, 2011. **286**(37): p. 32736-48.
17. Stawikowski, M.J., et al., *Glycosylation modulates melanoma cell alpha2beta1 and alpha3beta1 integrin interactions with type IV collagen.* J Biol Chem, 2014. **289**(31): p. 21591-604.
18. Pokidysheva, E., et al., *Posttranslational modifications in type I collagen from different tissues extracted from wild type and prolyl 3-hydroxylase 1 null mice.* J Biol Chem, 2013. **288**(34): p. 24742-52.
19. Magiorkinis, E., A. Beloukas, and A. Diamantis, *Scurvy: past, present and future.* Eur J Intern Med, 2011. **22**(2): p. 147-52.
20. Marini, J.C. and A.R. Blissett, *New genes in bone development: what's new in osteogenesis imperfecta.* J Clin Endocrinol Metab, 2013. **98**(8): p. 3095-103.
21. Malfait, F. and A. De Paepe, *The Ehlers-Danlos syndrome.* Adv Exp Med Biol, 2014. **802**: p. 129-43.
22. Ha-Vinh, R., et al., *Phenotypic and molecular characterization of Bruck syndrome (osteogenesis imperfecta with contractures of the large joints) caused by a recessive mutation in PLOD2.* Am J Med Genet A, 2004. **131**(2): p. 115-20.
23. Friedman, L., et al., *Prolyl 4-hydroxylase is required for viability and morphogenesis in Caenorhabditis elegans.* Proc Natl Acad Sci U S A, 2000. **97**(9): p. 4736-41.
24. Winter, A.D. and A.P. Page, *Prolyl 4-hydroxylase is an essential procollagen-modifying enzyme required for exoskeleton formation and the maintenance of body shape in the nematode Caenorhabditis elegans.* Mol Cell Biol, 2000. **20**(11): p. 4084-93.

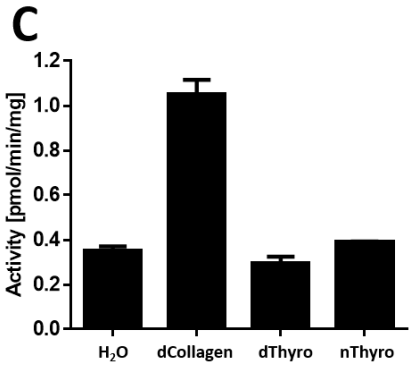
25. Holster, T., et al., *Loss of assembly of the main basement membrane collagen, type IV, but not fibril-forming collagens and embryonic death in collagen prolyl 4-hydroxylase I null mice.* J Biol Chem, 2007. **282**(4): p. 2512-9.
26. Heikkinen, J., et al., *Lysyl hydroxylase 3 is a multifunctional protein possessing collagen glucosyltransferase activity.* J Biol Chem, 2000. **275**(46): p. 36158-63.
27. Sricholpech, M., et al., *Lysyl hydroxylase 3 glucosylates galactosylhydroxylysine residues in type I collagen in osteoblast culture.* J Biol Chem, 2011. **286**(11): p. 8846-56.
28. Myllylä, R., et al., *Expanding the lysyl hydroxylase toolbox: new insights into the localization and activities of lysyl hydroxylase 3 (LH3).* J Cell Physiol, 2007. **212**(2): p. 323-9.
29. Rautavuoma, K., et al., *Characterization of three fragments that constitute the monomers of the human lysyl hydroxylase isoenzymes 1-3. The 30-kDa N-terminal fragment is not required for lysyl hydroxylase activity.* J Biol Chem, 2002. **277**(25): p. 23084-91.
30. Katzman, R.L., et al., *Isolation and structure determination of glucosylgalactosylhydroxylysine from sponge and sea anemone collagen.* Biochemistry, 1972. **11**(7): p. 1161-7.
31. Spiro, R.G., *The structure of the disaccharide unit of the renal glomerular basement membrane.* J Biol Chem, 1967. **242**(20): p. 4813-23.
32. Royce, P.M. and M.J. Barnes, *Comparative studies on collagen glycosylation in chick skin and bone.* Biochim Biophys Acta, 1977. **498**(1): p. 132-42.
33. Myllyla, R., L. Risteli, and K.I. Kivirikko, *Collagen glucosyltransferase. Partial purification and characterization of the enzyme from whole chick embryos and chick-embryo cartilage.* Eur J Biochem, 1976. **61**(1): p. 59-67.
34. Risteli, L., R. Myllyla, and K.I. Kivirikko, *Affinity chromatography of collagen glycosyltransferases on collagen linked to agarose.* Eur J Biochem, 1976. **67**(1): p. 197-202.
35. Trombetta, E.S. and A.J. Parodi, *Glycoprotein reglucosylation.* Methods, 2005. **35**(4): p. 328-37.
36. Shevchenko, A., et al., *In-gel digestion for mass spectrometric characterization of proteins and proteomes.* Nat Protoc, 2006. **1**(6): p. 2856-60.
37. Wisniewski, J.R., et al., *Universal sample preparation method for proteome analysis.* Nat Methods, 2009. **6**(5): p. 359-62.
38. Tenni, R., et al., *Hydroxylysine glycosides: preparation and analysis by reverse phase high performance liquid chromatography.* The Italian journal of biochemistry, 1984. **33**(2): p. 117-27.
39. Hülsmeier, A.J. and T. Hennet, *O-Linked glycosylation in Acanthamoeba polyphaga mimivirus.* Glycobiology, 2014. **24**(8): p. 703-14.
40. Myllyla, R., et al., *Isolation of collagen glucosyltransferase as a homogeneous protein from chick embryos.* Biochim Biophys Acta, 1977. **480**(1): p. 113-21.
41. Ishikawa, Y., et al., *Biochemical characterization of the prolyl 3-hydroxylase 1.cartilage-associated protein.cyclophilin B complex.* J Biol Chem, 2009. **284**(26): p. 17641-7.
42. Helenius, A. and M. Aebi, *Intracellular functions of N-linked glycans.* Science, 2001. **291**(5512): p. 2364-9.
43. Sousa, M. and A.J. Parodi, *The molecular basis for the recognition of misfolded glycoproteins by the UDP-Glc:glycoprotein glucosyltransferase.* EMBO J, 1995. **14**(17): p. 4196-203.
44. Tessier, D.C., et al., *Cloning and characterization of mammalian UDP-glucose glycoprotein: glucosyltransferase and the development of a specific substrate for this enzyme.* Glycobiology, 2000. **10**(4): p. 403-12.
45. Trombetta, S.E., M. Bosch, and A.J. Parodi, *Glucosylation of glycoproteins by mammalian, plant, fungal, and trypanosomatid protozoa microsomal membranes.* Biochemistry, 1989. **28**(20): p. 8108-16.
46. Arnold, S.M., et al., *Two homologues encoding human UDP-glucose:glycoprotein glucosyltransferase differ in mRNA expression and enzymatic activity.* Biochemistry, 2000. **39**(9): p. 2149-63.
47. Buzzi, L.I., et al., *The two Caenorhabditis elegans UDP-glucose:glycoprotein glucosyltransferase homologues have distinct biological functions.* PLoS One, 2011. **6**(11): p. e27025.

48. Arnold, S.M. and R.J. Kaufman, *The noncatalytic portion of human UDP-glucose: glycoprotein glucosyltransferase I confers UDP-glucose binding and transferase function to the catalytic domain.* J Biol Chem, 2003. **278**(44): p. 43320-8.
49. Takeda, Y., et al., *Both isoforms of human UDP-glucose:glycoprotein glucosyltransferase are enzymatically active.* Glycobiology, 2014. **24**(4): p. 344-50.
50. Xu, C. and D.T. Ng, *O-mannosylation: The other glycan player of ER quality control.* Semin Cell Dev Biol, 2015.
51. Smith, M.H., H.L. Ploegh, and J.S. Weissman, *Road to ruin: targeting proteins for degradation in the endoplasmic reticulum.* Science, 2011. **334**(6059): p. 1086-90.
52. Ushioda, R., J. Hoseki, and K. Nagata, *Glycosylation-independent ERAD pathway serves as a backup system under ER stress.* Mol Biol Cell, 2013. **24**(20): p. 3155-63.
53. Acar, M., et al., *Rumi is a CAP10 domain glycosyltransferase that modifies Notch and is required for Notch signaling.* Cell, 2008. **132**(2): p. 247-58.
54. Sethi, M.K., et al., *Molecular cloning of a xylosyltransferase that transfers the second xylose to O-glucosylated epidermal growth factor repeats of notch.* J Biol Chem, 2012. **287**(4): p. 2739-48.

FIGURES AND TABLES







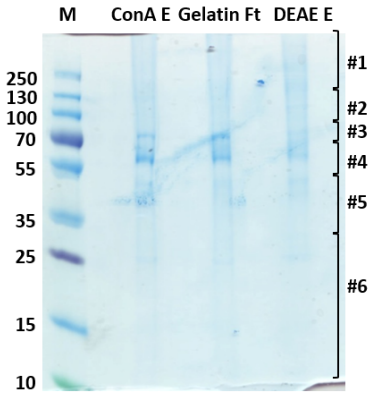
D

Purification of ColGlcT from chicken embryo homogenates

Step	Protein	Activity		Purification
	mg/ml	pmol/min/mg	cpm/mg	
Homogenate (Hg)	21.47	0.010	928	
$(NH_4)_2SO_4$	34.06	0.012	1'143	1.2
ConA-Seph	3.13	0.264	24'736	26.4
DEAE-Seph	0.66	1.052	98'690	**105.2**

Figure 1| **Enrichment of ColGlcT activity from chicken embryo.** After $(NH_4)_2SO_4$ precipitation, proteins were subjected to two-step chromatography, namely ConA-sepharose and anion exchange DEAE-sepharose. **(A)** ColGlcT activity was determined after every step by measuring the [^{14}C]-labelled glucose incorporation to denatured collagen. **(B)** The specific ColGlcT activity was calculated based on the amount of donor substrate and protein concentration in the sample. **(C)** The eluted fraction B5 was tested for different acceptor substrates with the highest activity for denatured collagen (dcollagen). **(D)** The ColGlcT activity could be enriched by 105-fold after purification from chicken embryo homogenate. Ld = load, SN = Supernatant, Ft = flowthrough, W = wash, E = elution, nThyro = native thyroglobulin, dThyro = denatured thyroglobulin. Mean ± SEM, n=2. Representative of 4 independent experiments.

A



B

LC-MS/MS protein identification

Gel region	ConA	Gelatin	DEAE
ingel 1	201	197	234
ingel 2	199	256	297
ingel 3	144	180	207
ingel 4	159	186	167
ingel 5	197	185	153
ingel 6	237	268	289
	1137	1272	1347
whole proteome	**701**	**822**	**933**
in all 3		437	
enriched in DEAE over Gela/ConA (>1)		181	

Figure 2| **Protein identification of ColGlcT active fractions.** SDS-PAGE of chromatographed fractions after ConA, gelatin, and anion exchange chromatography. # indicates region cut out of the gel and used for peptide sample preparation for LC-MS/MS. (B) LC-MS/MS protein identification for indicated region.

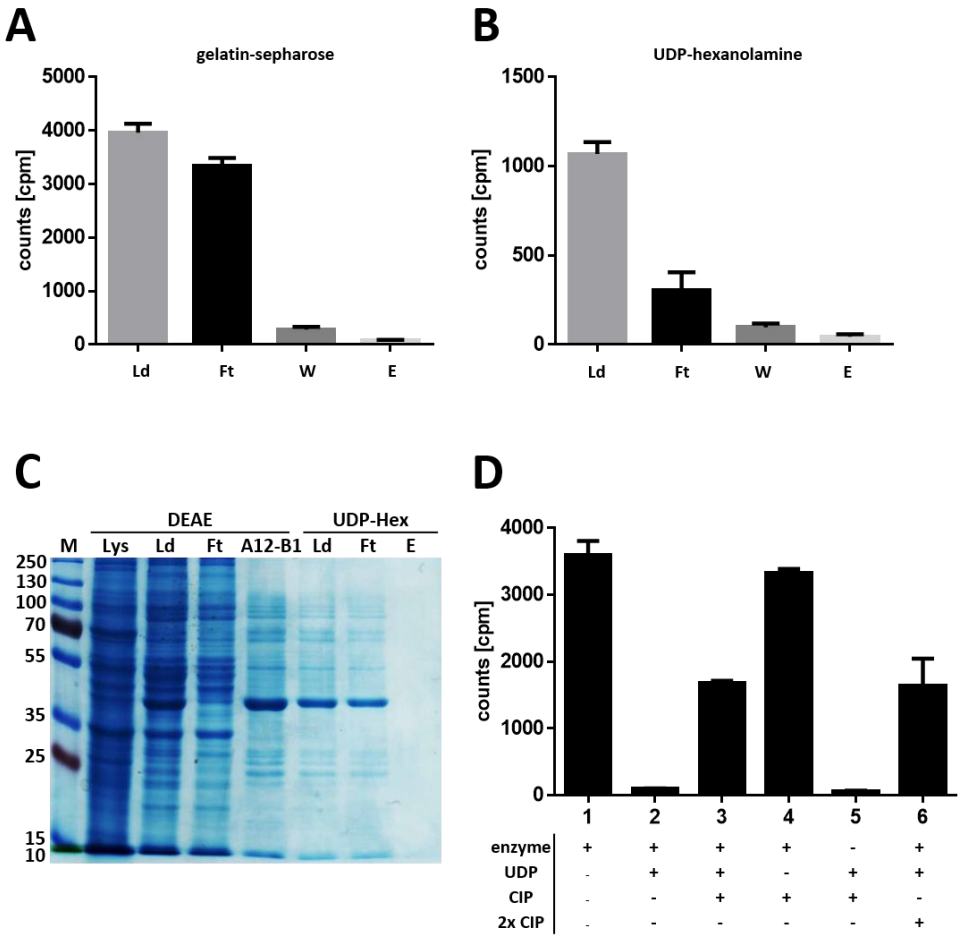


Figure 3| **No elution of ColGlcT from gelatin-sepharose or UDP-hexanolamine-agarose.** Total cell lysates from chicken embryo homogenates or human fibroblasts have been prefractionated with anion exchange chromatography prior to run over **(A)** gelatin-sepharose or **(B)** UDP-hexanolamine-agarose affinity chromatography, respectively. After gelatin-sepharose chromatography the ColGlcT activity was measured in the flowthrough (Ft) fraction indicating no binding. No ColGlcT activity could be measured after UDP-hexanolamine chromatography after elution with up to 10mM UDP. **(C)** SDS-PAGE of the eluted fraction E did not show any eluted proteins. **(D)** The inhibiting effect of UDP in the ColGlcT assay could be reduced by 50% when adding calf intestine phosphatase (CIP) to the reaction (#3). Doubling the amount of CIP does not reduce inhibition anyfurther (#6). Ld = load, Ft = flowthrough, W = wash, E = elution. Mean ± SEM, n=2.

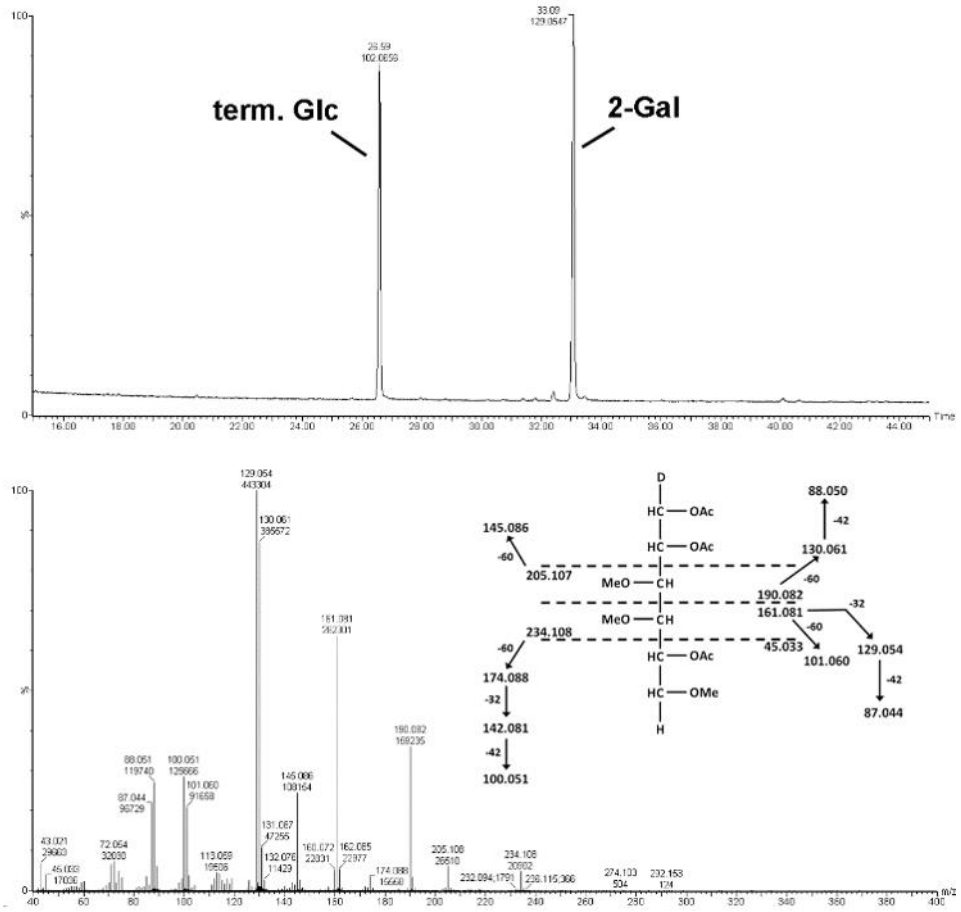


Figure 4| **Methylation linkage analysis of purified glucose-galactose-hydroxylysine from sponge**. Partially methylated alditol acetates were prepared from glucose-galactose-hydroxylysine and subjected to GC-MS analysis. Panel A shows the total ion chromatogram of the peaks corresponding to the terminal glucose and the 2-substituted galactose. Panel B shows the annotated electron impact ionization mass spectrum of 3,4,6-trimethyl-1,2,5-acetyl-galactitol derived from the 2-substituted galactose eluting at 33 min.

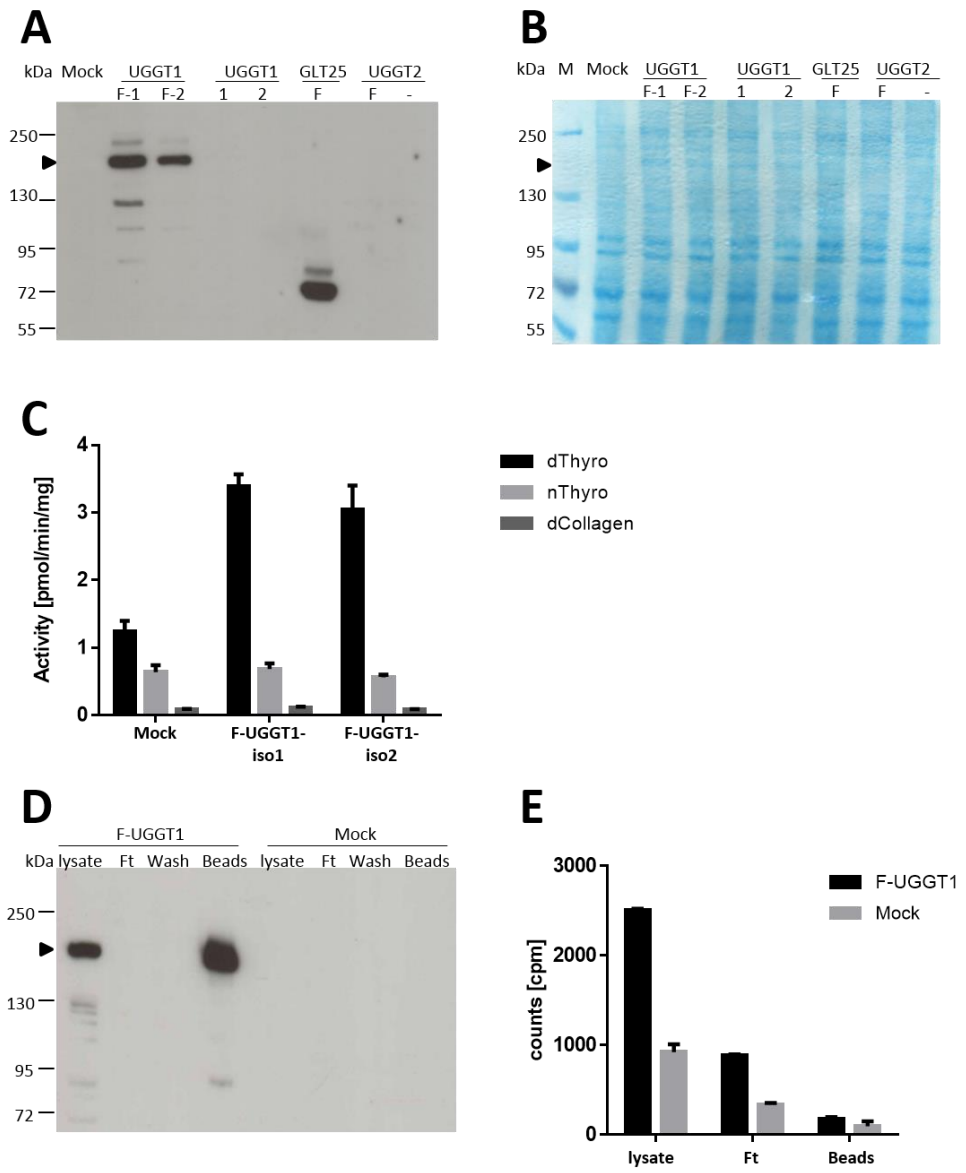


Figure 5| **Glucosyltransferase activity of Flag-UGGT1 on heat denatured thyroglobulin but not on collagen. (A, B)** Flag-tagged UGGT1 isoform 1 and 2 were expressed using the baculoviral/insect-cell expression system. **(C)** Activity of total cell lysates was determined when incubated with heat denatured thyroglobulin, native thyroglobulin or heat denatured collagen for 15 min or 180 min, respectively. **(D)** Flag-purification of F-UGGT1 and stained with anti-Flag antibody in western blot. **(E)** No activity could be measured of Flag-purified F-UGGT1 on denatured thyroglobulin. Mock = empty vector control. Ft = flowthrough. Beads = Enzyme-bound bead fraction. Mean ± SEM, n=3.

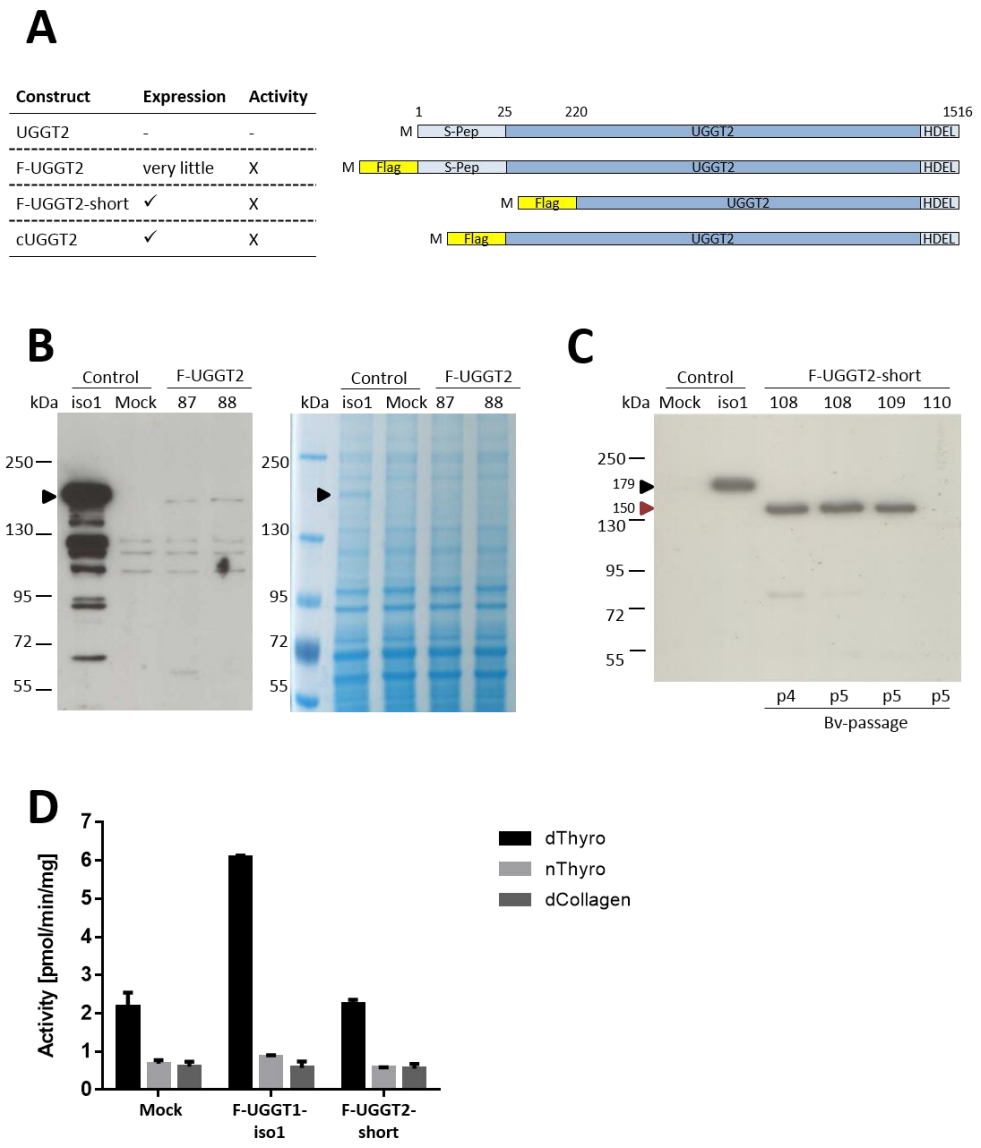


Figure 6| **F-UGGT2 constructs**. Panel (A) represents the different UGGT constructs expressed in the baculoviral expression system. The Flag-tagged full-length UGGT2 (F-UGGT2) including the N-terminal signal sequence (S-Pep) was only expressed weakly even though the bacuviral infections were good (B). An N-terminal truncated version F-UGGT2-short could be expressed but has no glucosylation activity compared to UGGT1 (C-D). The forth construct (cUGGT2) is characterized by the removal of only the 25 amino acid S-Pep and is described in Figure 7.

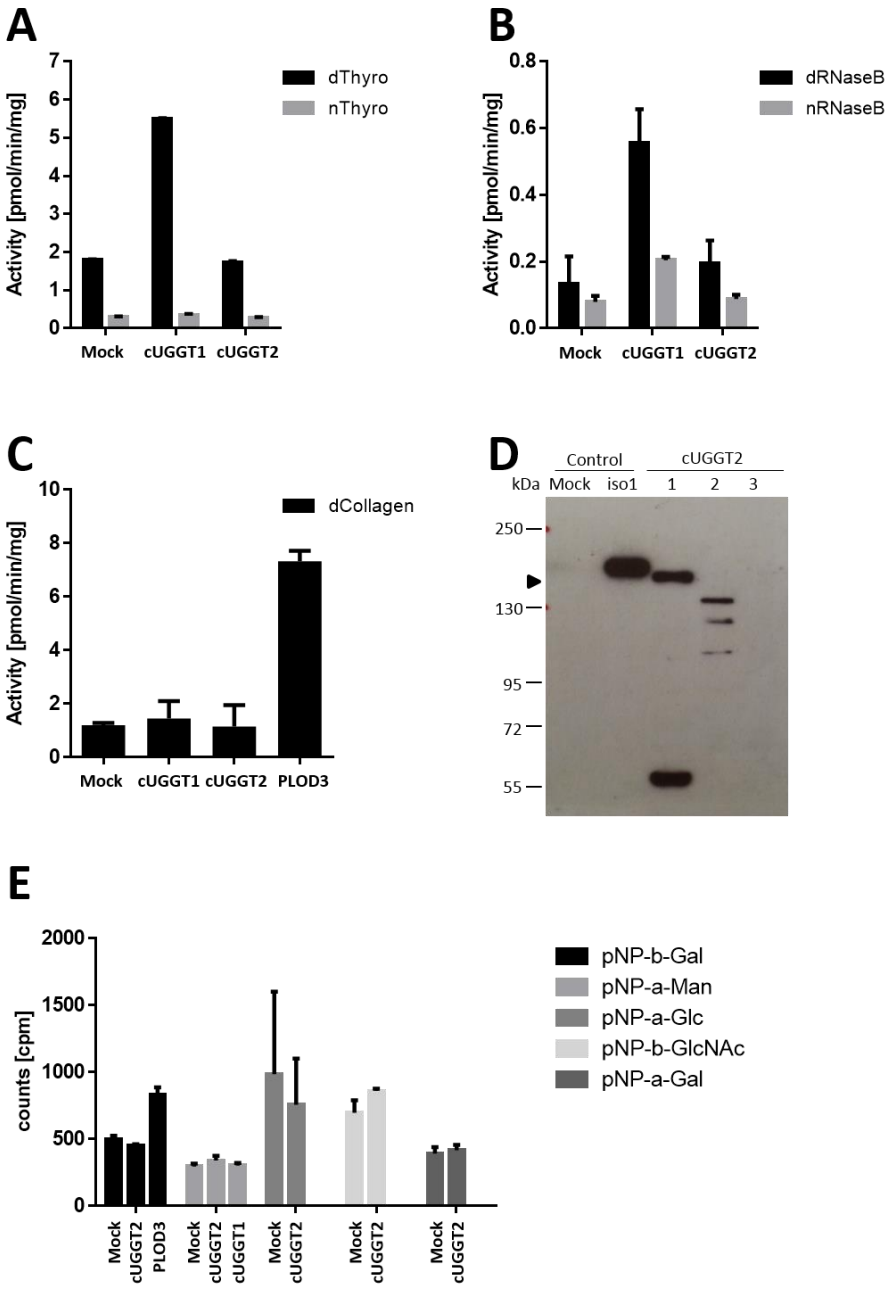


Figure 7| **No activity of UGGT2 towards thyroglobulin, RNaseB and collagen.** Baculoviral expressed cUGGT2 (c = without signal peptide) is not active on **(A)** denatured thyroglobulin, **(B)** denatured RNaseB and **(C)** denatured collagen. Mean ± SEM, n=3 **(D)** Expression of UGGT1 and UGGT2 without signal peptide (cUGGT1 and cUGGT2) visualized by anti-Flag western blot. **(E)** No activity of cUGGT2 on 5 different pNP-sugars. Mean ± SEM, n=2

Table 1. LC-MS/MS protein identification from SDS-Page after anion exchange (DEAE).

#	Accesion	Description	score[a]	emPAI[b]	MW[c]
R1		Endoplasmin	550	0.75	91.7
		Tenascin	411	0.27	204
		Collagen alpha-1(XIV) chain	357	0.31	203.7
	E1BVX3	UDP-glucose:glycoprotein glucosyltransferase 1	196	0.16	177.8
		Glycogen [starch] synthase, muscle	155	0.26	84.8
	F1NX83	Glycogen debranching enzyme	64	0.04	176.8
		Neutral alpha-glucosidase AB	52	0.06	107.3
		UDP-glucose:glycoprotein glucosyltransferase 2	47	0.02	175.3
R2		Endoplasmin	1593	8.13	91.7
		Heat shock protein HSP 90-alpha	443	0.91	84.4
		Heat shock cognate 71 kDa protein	677	1.26	71.0
		78 kDa glucose-regulated protein	292	0.70	72.5
		Glycogen [starch] synthase, muscle	210	0.35	84.8
		Collagen alpha-1(I) chain, fragment	134	0.12	138.7
	P24802	PLOD1, lysyl-hydroxylase 1 (LH1)	175	0.26	84.8
		PLOD2, lysyl-hydroxylase 2 (LH2)	57	0.04	85.1
		Collagen triple helix repeat-cont. Protein 1	44	0.26	27.0
R3		Heat shock cognate 71 kDa protein	1106	5.09	71.1
		Heat shock 70 kDa protein	608	1.50	69.9
		78 kDa glucose-regulated protein	477	1.03	72.5
	P24802	PLOD1, lysyl-hydroxylase 1 (LH1)	175	0.21	84.8
		Protocadherin-10	89	0.12	114.2
		Beta-1,4-galactosyltransferase 1	28	0.07	44.8
R4		Protein disulfide-isomerase A3	234	1.34	56.5
		Protein disulfide-isomerase (P4HB)	184	0.95	57.8
		Sarcalumenin	125	0.34	54.6
	F1N9J1	Glycogenin-1	42	0.09	37.6
		Glucosidase 2 subunit beta	40	0.11	59.7
R6		ERp29 fragment	35	0.13	25.5
		Uncharacterized protein C11orf73 homolog	27	0.15	21.9

a = Mascot score for protein identification probability. Score > 25 indicates positive identification with p = 0.05
b = emPAI (exponentially modified protein abundance index)describes the protein composition in the sample
c = molecular weight

Table 2. Enrichment of proteins after anion exchange (DEAE). The protein emPAI values, indicating the protein abundance in the whole proteome of the eluted DEAE fraction, were compared to the emPAI values after ConA and gelatin affinity chromatography. A selection of proteins with enrichment values above 1, indicating specific protein enrichment, are shown in this table.

Accesion[a]	Description[b]	ConA		Gelatin		DEAE		Enrichment[c]
		score[d]	emPAI[e]	score	emPAI	score	emPAI	
P00760	TRYP_BOVIN TRYPSINOGEN	2336	22.39	1803	21.05	2575	27.47	**1.1**
F1NM70	Follistatin-related protein 1	115	0.41	172	0.41	196	1.44	**3.0**
E1C707	Inosine-5'-monophosphate dehydrogenase 2	29	0.06	25	0.06	453	1.7	**24.3**
F1P4H4	Thioredoxin domain-containing protein 5	62	0.31	84	0.4	163	0.61	**1.5**
F1NWB7	Endoplasmin	1754	4.66	1684	5.27	3191	16.29	**2.8**
E1BVX3	UDP-glucose:glycoprotein glucosyltransferase 1	252	0.23	310	0.29	826	0.89	**2.9**
E1C6L8	Cell surface glycoprotein MUC18	19	0.13	88	0.28	87	0.4	**1.7**
F1NKJ6	Inactive serine protease PAMR1	72	0.13	129	0.17	343	0.79	**4.5**
F1P2Z2	VPS10 domain-containing receptor SorCS1	191	0.21	142	0.15	530	0.49	**2.4**
F1NWJ4	Protocadherin-10	138	0.16	227	0.26	347	0.4	**1.6**
Q8AV57	Protein sidekick-2	194	0.24	244	0.27	638	0.72	**2.4**
F1NMF6	Procollagen-lysine,2-oxoglutarate 5-dioxygenase 1	117	0.16	150	0.08	428	0.52	**3.8**
Q90593	78 kDa glucose-regulated protein	491	1.32	826	2.05	1061	2.59	**1.3**
F1NL92	Cartilage matrix protein (Fragment)	50	0.12	22	0.06	244	0.58	**5.6**
H9KZZ3	Lysyl oxidase homolog 3	32	0.04	44	0.09	138	0.23	**3.0**

a = protein annotation from the Gallus Search Database (FGCZ)

b = protein description after blasting the accesion against the human protein database (NCBI)

c = the enrichment was calculated according the emPAI valuees of the whole identified proteome of the DEAE proteome over the ConA/Gelatin

d = Mascot score for protein identification probability. Score > 25 indicates positive identification with p = 0.05

e = emPAI (exponentially modified protein abundance index)describes the protein composition in the sample

Table 3. LC-MS/MS protein identification of fraction B5 from Figure 1. The fraction B5 was subjected to three different alkylation protocols prior to shoot on LC-MS/MS. RapiGest is a molecule that helps solubilizing proteins and is precipitated prior to LC. FASP is a filter aided sample preparation protocol also supporting to purify membrane associated proteins. Ingel refers to the standard protocol after SDS-Page. The table shows a selection of proteins identified with the three different protocols.

Accesion[a]	Description[b]	RapiGest		FASP		Ingel	
		score[c]	emPAI[d]	score	emPAI	score	emPAI
F1N9J1	Glycogenin-1	197	0.41	1234	6.7	1066	3.22
E1C707	Inosine-5'-monophosphate dehydrogenase 2	209	0.33	419	0.67	467	0.98
F1NJW3	Uncharacterized protein (Fragment)	236	1.13	224	1.13	152	0.74
E1C303	1,4-alpha-glucan-branching enzyme	215	0.15	405	0.68	190	0.2
E1BVX3	UDP-glucose:glycoprotein glucosyltransferase 1	63	0.04	192	0.14	517	0.26
F1NX83	Glycogen debranching enzyme	159	0.18	229	0.14	214	0.2
F1P574	Mannose-1-phosphate guanyltransferase beta	26	0.1	87	0.48	24	0.1
F1P2Z2	VPS10 domain-containing receptor SorCS1	116	0.06	65	0.03	266	0.15
E1BQM1	Cysteine-rich with EGF-like domain protein 2	361	0.99	61	0.08	55	0.17
F1P4N9	Periostin	594	0.78	861	1.26	1304	1.9
P24802	PLOD1	84	0.04			262	0.16
E1BZL6	Chondroitin sulfate synthase 3	43	0.06				
Q5F3N9	Uncharacterized protein C6orf106 homolog			61	0.1		
Q5ZKG8	Golgi to ER traffic protein 4 homolog			44	0.1		
E1C4S6	Uncharacterized protein			19	0.5		
	Protein disulfide-isomerases	A3, A4,A6, P4Hb, with all three protocols					
	ER-chaperones	Endoplasmin, GRP78, with all three protocols					

a = protein annotation from the Gallus Search Database (FGCZ)
b = protein description after blasting the accesion against the human protein database (NCBI)
c = Mascot score for protein identification probability. Score > 25 indicates positive identification with p = 0.05
d = emPAI (exponentially modified protein abundance index)describes the protein composition in the sample

2) Affinity purification with PLOD3

Until now the collagen lysyl hydroxylase 3 (LH3) encoded by *PLOD3* is considered to glucosylate the galactosyl-hydroxylysine residues from collagen [1, 2]. Compared to the ColGalT enzyme GLT25D1, the conversion rates for PLOD3 activity are low and its sole contribution to ColGlcT activity remains questionable. Since we found that overexpression of PLOD3 appears to have some little activity for collagen glucosylation (Figure 6A) the question arises whether PLOD3 itself is sufficient for the collagen glucosylation or whether PLOD3 is part of an enzymatic complex together with the ColGlcT which then glucosylates collagen. Other collagen modifying enzymes also act in protein complexes like the P3H complex [3, 4], suggesting that overexpression of one of the complex members could increase activity levels of the complex. Considering ColGlcT being part of the PLOD3 forming complex, we co-immunoprecipitated differentially-tagged PLOD3 from PLOD3 overexpressed cell lysates to identify the potential interacting partners of PLOD3. We used two different tagging strategies. On the one hand, we fused a human Fc-tag at the C-terminus of PLOD3 which was then immunoprecipitated with Protein A (Figure 8A). On the other hand, we cloned the tandem affinity purification TAP-tag at the N-terminus of PLOD3, ensuing the TAP-PLOD3 construct which includes a two-step affinity purification protocol using IgG and streptavidin affinity resins (Figure 9A). Both strategies were applied independently and we searched the PLOD3 purified fractions after mass spectrometric protein identification for glycosyltransferases.

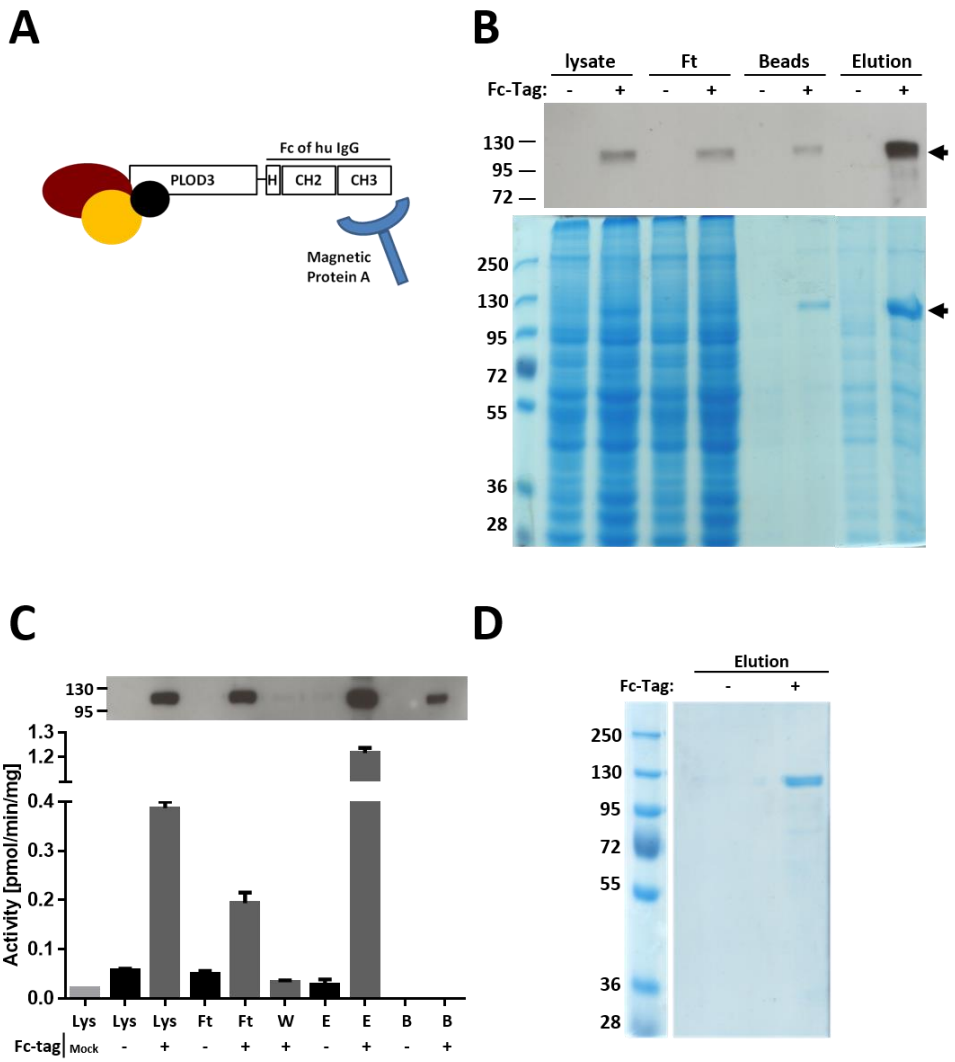


Figure 8|**ColGlcT affinity purification with Fc-tagged LH3. A)** Schematic representation of the C-terminal human Fc-tagged LH3 protein encoded by PLOD3 and its interacting partners shown in colored rounds. **B)** SDS-PAGE and anti-human IgG Western blot of the Fc-purified LH3 **C)** ColGlcT activity was determined after every purification step for the Fc-tagged LH3 (PLOD3-Fc*) and the untagged LH3 protein (PLOD3). The top panel shows the anti-human IgG Western blot for the corresponding fractions. Ft: flow through fraction, W: wash fraction, Elu: eluted fraction. **D)** SDS-PAGE of the eluted fractions for the untagged (-) and tagged (+) LH3 upon intensified washing steps.

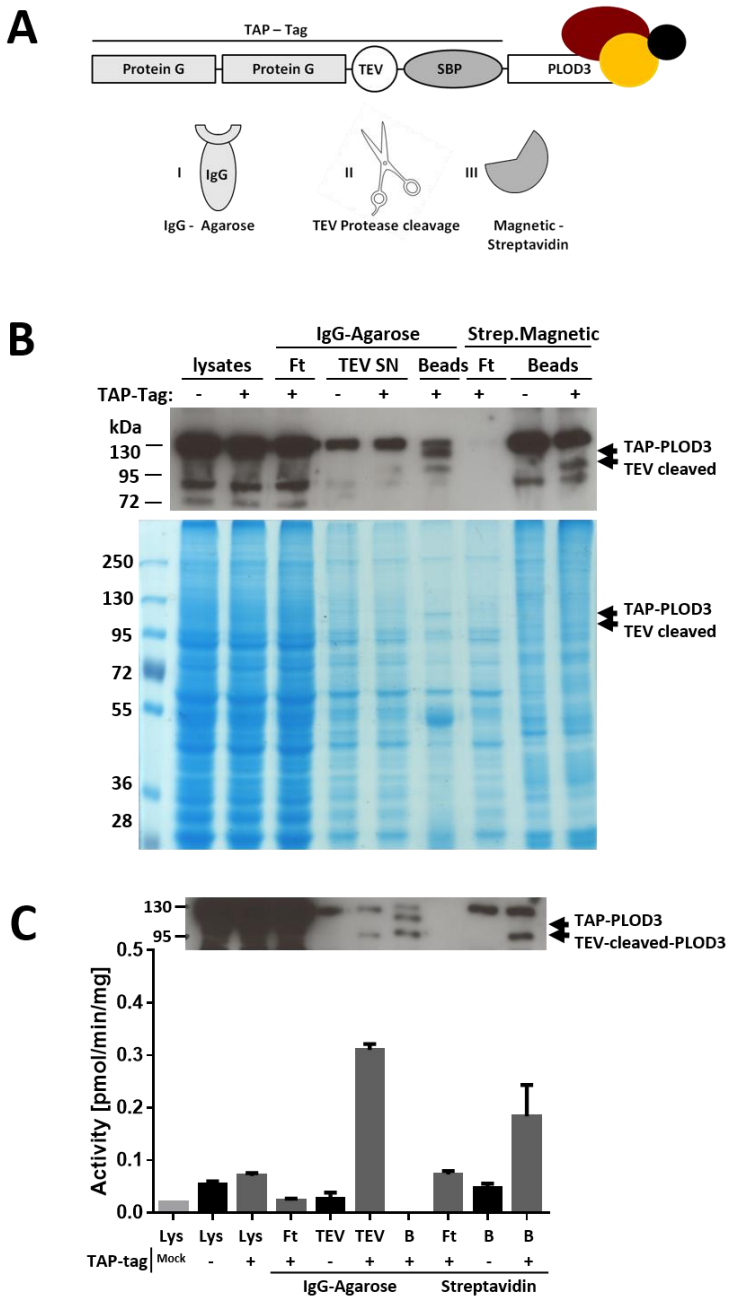


Figure 9| **ColGlcT affinity purification with TAP-tagged LH3. A)** Schematic representation of the N-terminal TAP-tagged LH3 protein encoded by PLOD3 and its interacting partners shown in colored rounds. **B)** SDS-PAGE and Streptavidin-POD Western blot of the TAP-purified LH3. TAP-PLOD3 indicates the 108 kDa full construct and the TEV cleaved construct is at 92 kDa. **C)** ColGlcT activity was determined after every purification step. (+) indicates TAP-tagged LH3 input and (-) indicates untagged LH3 input. The top panel shows the Streptavidin-POD Western blot for the corresponding fractions. Ft: flow through fraction, W: wash fraction, B: protein-bead complex.

LC-MS/MS protein identification of the PLOD3 affinity purified fractions

UDP-glucose:glycoprotein glucosyltransferase 2 (UGGT2) is the only glycosyltransferase that could be identified from PLOD3 affinity purified fractions (Table 2). Interestingly, UGGT2 has been found after application of both independent tagging strategies. By determining the specific ColGlcT activity after each purification step we measured activity in the purified fractions containing PLOD3 and UGGT2 (Figure 2B, C and 3B, C). However, the identification of UGGT2 could not be verified in a second experiment and considering the vast amount of proteins in the PLOD3 affinity purified fraction, the identification of UGGT2 most possibly does not come from specific interaction with PLOD3. No glycosyltransferases were detected repeatedly with either of the tagged constructs. We found three ER chaperone proteins which could be identified multiple times, the 78-kDa glucose regulated protein (GRP78), luminal binding protein 4 (BIP4) and heat shock 70 kDa cognate (HSC-70) (Table 2). We did not further investigate their interaction with PLOD3 since we couldn't detect a single band on SDS-PAGE, like for PLOD3, indicating not a strong binding. Most probably the chaperones are not specifically purified upon PLOD3 interaction but rather contaminating the sample due to their high abundance. In order to distinguish whether the activity we measured results from purified PLOD3 itself and to identify strong interacting partners we next repeated the experiment with more intensive washing steps to reach higher purity.

Table 2| **LC-MS/MS protein identification of the PLOD3 affinity purified fractions**

Accession	Description	1st	2nd
		393	90
sp|O60568|PLOD3_HUMAN	**Procollagen-lysine,2-oxoglutarate 5-dioxygenase 3**	**x**	**x**
sp|P01857|IGHG1_HUMAN	Ig gamma-1 chain C region	x	x
sp|P81605|DCD_HUMAN	Dermcidin	x	x
sp|P29844|HSP7C_DROME	Heat shock 70 kDa protein cognate 3	x	x
sp|P06761|GRP78_RAT	**78 kDa glucose-regulated protein**	**x**	**x**
sp|Q03684|BIP4_TOBAC	Luminal-binding protein 4	x	x
Q9NVE4|CCD87_HUMAN	Coiled-coil domain-containing protein 87	x	
Q3SH73|G6PI2_THIDA	Glucose-6-phosphate isomerase 2	x	
Q9NYU1|UGGG2_HUMAN	**UDP-glucose:glycoprotein glucosyltransferase 2**	**x**	
sp|Q00733|VP80_NPVAC	Capsid protein p80		x
sp|Q68DL7|CR063_HUMAN	**Uncharacterized protein C18orf63**	**x**	
sp|Q58FG1|HS904	Putative heat shock protein HSP 90-alpha A4	x	
sp|Q9XTL9|PYG_DROME	Glycogen phosphorylase	x	
sp|A5FG96|GLGA_FLAJ1	Glycogen synthase	x	
sp|P18569|UDPE_NPVAC	Ecdysteroid UDP-glucosyltransferase	x	
sp|P0C1H9|PPIB1_RHIO9	Peptidyl-prolyl cis-trans isomerase B1	x	
sp|Q14117|DPYS_HUMAN	Dihydropyrimidinase	x	
sp|Q9M069|E137_ARATH	Glucan endo-1,3-beta-glucosidase 7	x	
sp|B2FLK4|GH109_STRMK	Glycosyl hydrolase family 109 protein	x	
sp|Q5RBQ2|GINM1_PONAB	Glycoprotein integral membrane protein	x	

Purified PLOD3 glucosylates heat denatured collagen

Purification of Fc-tagged PLOD3 to higher purity increases the specific ColGlcT activity. Higher purity was achieved by introducing more washing steps from Protein A coupled magnetic Dynabeads. On SDS-PAGE only the purified PLOD3 protein could be detected and mass spectrometric analysis of the same fraction revealed no glycosyltransferase being present in that fraction (Figure 8D). According to Western blot analysis we could enrich PLOD3 in the purified fraction but when looking at the raw counts, the enzymatic activity of PLOD3 did not increase (Figure 8C). The acidic elution with pH 5 might kill some of the enzymes activity even though the eluted fraction is neutralized immediately. The higher enzymatic activity in the lysate could also indicate for a second enzyme contributing for ColGlcT activity.

3) Candidate search from database – expression and activity

Uncharacterized glycosyltransferases from CAZy families 8, 61 and 90 do not glucosylate collagen

We searched the Carbohydrate Active enZYmes database (CAZy) for possible ColGlcT candidate genes according to families related to predicted α-glucosyltransferases and selected genes with uncharacterized functions (Figure 10). The CAZy-database comprises glycosyltransferases, which are grouped into families based on structural relations and homology of the catalytic modules. The glycosyltransferase families GT8, GT24 and GT90 encompass enzymes with retaining α-glucosyltransferases. The GT8 is a large family including the human glycogenin-1 and -2, LARGE 1 and 2, GLT8D1-4 and XXylT1. Unlike glycogenin and LARGE which are very well described

GT-fold	GT-family *	Human family members	Unknown, uncharacterized members
GTA – retaining	6	BGAT(ABO), GBGT1, GLT6D1	GLT6D1
	8	GYG1, GYG2, LARGE1, LARGE2, GLT8D1 – GLT8D4, XXYLT	GLT8D1,GLT8D2
	24	UGGT1, UGGT2	UGGT2
	27	GALNT1 – 15	
	32	A4GNT, A4GALT	
	64	EXT1, EXT2, Extlike1 – Extlike3	Extlike3
GTB – retaining	3	GYS1, GYS2	
	4	ALG2, ALG11, PIGA, GTDC1	GTDC1
	35	PIGM	
GT – unknown	22, 39, 50, 57, 58, 59, 66, 76	All described human GT-families with unknown GT-folds use Dol-P-sugars as donor	
GT – unclassified	nc	PLOD3	

* GT-families containing human sequences according CAZy.org

Figure 10| **Small selection of glycosyltransferases retrieved from the CAZy database.** The CAZy GT-families were selected based on the appearance of human sequences within the retaining GT-families. Members of the selected families that are not adequately described or characterized were chosen to clone into the baculovirus/expression system.

glucosyltransferases, only little is known about the GLT8D proteins. GLT8D3 and D4 are α-1,3 xylosyltransferases elongating core-glucosylation on Notch EGF-repeats [5]. We cloned GLT8D1 and D2 into our baculovirus expression system and tested for ColGlcT activity. Both enzymes do not glucosylate denatured collagen (Figure 11). When the enzymes activity was assayed on *p*-Nitrophenyl-β-galactopyranoside (*p*NP-βGal), as an alternative acceptor substrate, we measured activity for GLT8D2 (Figure 11B). But, upon repeating experiments, we could not verify GLT8D2 activity towards *p*NP-βGal. With the broad existence of collagen and its conserved PTM including glycosylation, the glycosylating enzymes are thought to be conserved as well among the animal kingdom. By blasting the GLT8D family members, we could not find any homologues proteins in the nematode *Caenorhabditis elegans* or the sponge *Amphimedon queenslandica* proteome, indicating that GLT8D1 or GLT8D2 are not the ColGlcT (Figures 14 and 15).

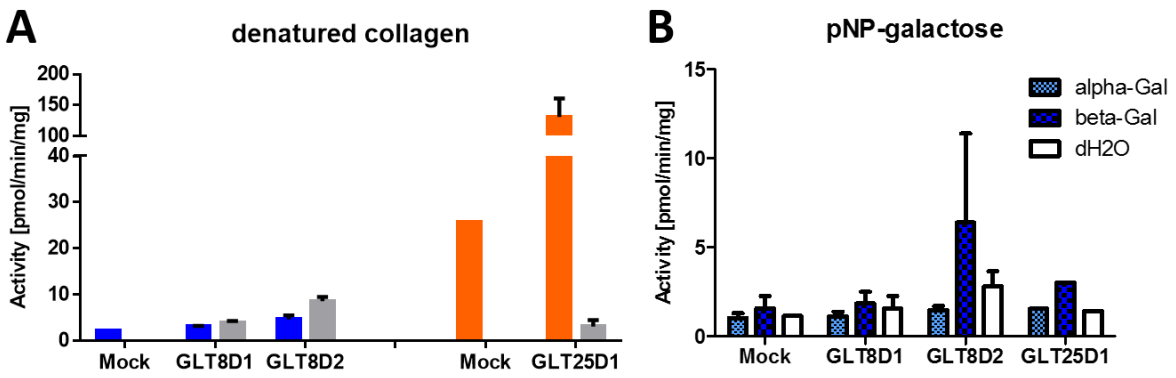
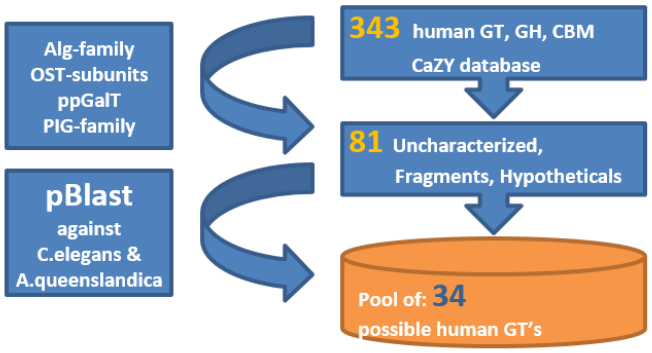


Figure 11| **ColGlcT activity of GLT8D1 and GLT8D2 on denatured collagen or *p*NP-galactose.** The human cDNA from *GLT8D1* and *GLT8D2* was cloned into the baculovirus/expression system and ColGlcT activity was addressed either with (A) denatured bovine type I collagen or (B) *p*NP-alpha-Gal or *p*NP-beta-Gal as acceptor substrates. In panel (A) GLT25D1 was used as an expression control with [^{14}C]-Gal as donor substrate (orange). Blue indicates [^{14}C]-Glc as donor substrate and grey indicates the control without acceptor substrate. n = 2, Mean ± maximum deviation from mean.

Broader analysis of the human list from the CAZy-database revealed 343 glycosyltransferases (GT), glycoside hydrolases (GH) and carbohydrate-binding modules (CBM) (Figure 12). To narrow down the list, all well described glycosyltransferases were removed from the list. Enzymes involved in the N-glycosylation pathway like the Alg family members and the oligosaccharyltransferase subunits (OST) were eliminated resulting in a list of 81 mainly uncharacterized or unknown glycosytransferases. We performed extensive blast searches against the complete proteomes of the nematode *C. elegans* and the sponge *A. queenslandica* which both contain collagen proteins and its modifying enzymes. All proteins for which we couldn't identify a homologue were removed from the list reducing it to 34 candidate proteins. Among the candidates



Accesion	Protein Name	Protein Info	Family
CAI30569.1	Aer61 (DKFZp686M05189)	uncharacterized glycosyltransferase AER61	GT61
AAA58642.1	branching-enzyme (Gbe1)	1,4-alpha-glucan-branching enzyme	CBM48,GH13
AAH18734.1	CGI-14 protein	putative N-acetylglucosamine-6-phosphate deacetylase	CE9
XP_069745.1	ENSP00000194108 (fragment)	glucosamine--fructose-6-phosphate aminotransferase	GH56
AAQ16408.1	Flj11753 (GTDC1)	glycosyltransferase-like domain-containing protein 1	GT4
CAE46499.1	Fut10		GT10
AAH36037.1	Fut11	alpha-(1,3)-fucosyltransferase 10	GT10
XP_069244.1	Gcnt6 (LOC135239)	beta-1,3-galactosyl-O-glycosyl-glycoprotein beta-1,6-N-acetylglucosaminyltransferase 3	GT14
AAH02617.1	glycogen synthase (Gys1;Gys)	glycogen [starch] synthase, muscle	GT3
CAA05859.1	glycogen synthase (Gys2)		GT3
AAH00033.1	glycogenin (Gyg;Gyg1;GYG1)	glycogenin-1	GT8
AAB84377.1	glycogenin 2 (Gyg2)		GT8
CAC27251.1	LOC129530	deleted in malignant brain tumors 1 protein	GH23
XP_071213.1	LOC138986 hypothetical protein XP_071213	hypothetical protein LOC100641567 decaprenyl-diphosphate synthase subunit 1	GH23
BAC11247.1	LOC89944 / BC008326 (Glb1l2)	beta-galactosidase-1-like protein 2	GH35
CAC51167.1	MGC10771	beta-galactosidase	GH35
AAQ88849.1	MGC4655 galactosyltransferase	beta-1,3-galactosyltransferase 6	GT31
AAP56253.1	myelodysplastic syndromes relative protein (Ktelc1;MdsrP)	protein O-glucosyltransferase 1	GT90
BAG62808.1	ORF (possible fragment)	galactoside 2-alpha-L-fucosyltransferase 2	GT11
AAA86946.1	oviduct glycoprotein (Ovgp1;Ogp)		GH18
XP_051350.1	pseudogene (fragment)	alpha-glucosidase	GH31
AAD45831.1	UDP-Glc: 5-(D-galactosyloxy)-L-lysine-procollagen α-glucosyltransferase	Plod3 / LH3 lysyl hydroxylase 3	GTnc
AAH38711.1	UDP-Glc: ceramide glucosyltransferase (UGCG)	UGCG	GT21
AAH41098.1	UDP-Glc: glycoprotein α-glucosyltransferase 1	HUGT1	GT24
AAF66233.1	UDP-Glc: glycoprotein α-glucosyltransferase 2	HUGT2	GT24
AAC02898.1	UDP-GlcNAc: α-1,4-N-acetylglucosaminyltransferase (Exotosin-like 2)	exostosin-like 2	GT64
AAB62283.1	UDP-GlcUA:β-glucuronyltransferase / UDP-GlcNAc: α-N-acetylglucosaminyltransferase	exostosin-1b	GT47,GT64
AAB07008.1	UDP-GlcUA:β-glucuronyltransferase / UDP-GlcNAc: α-N-acetylglucosaminyltransferas	exostosin-2	GT47,GT64
BAD18495.1	unnamed protein product	lysosomal alpha-glucosidase	GH31
CAD35071.1	unnamed protein product (contains ENSP00000171744)	chondroitin sulfate synthase 1	GT31
BAD18395.1	unnamed protein product (contains dJ1153D9.2)	beta-1,3-galactosyl-O-glycosyl-glycoprotein beta-1,6-N-acetylglucosaminyltransferase 3	GT14
CAC34689.1	unnamed protein product (contains FLJ00228)	hypothetical protein LOC100638402	GH65
BAC11155.1	unnamed protein product (fragment)	beta-galactosidase-1	GH35
AAO37647.1	UDP-Glc: TSR-fucose β-1,3-glucosyltransferase	hypothetical protein LOC100631565, B3GTL	GT31

Figure 12| **List of 34 sequences from the human CAZy-database** which contains unknown or uncharacterized glycosyltransferases with homologues in the *C.elegans* or *A.queenslandica* genomes.

is the uncharacterized glycosyltransferase AER61 from the GT61 family. The GT61 family contains 1,2-xylosyltransferases which are similar to glucosyltransferases. The two human proteins belonging to this family were both non-characterized. We cloned AER61 into the baculovirus

expression system and analyzed for collagen glucosyltransferase. No ColGlcT activity was measured for AER61 (Figure 13A). By now, AER61 has been described as the ER- O-GlcNAc transferase (EOGT) [6].

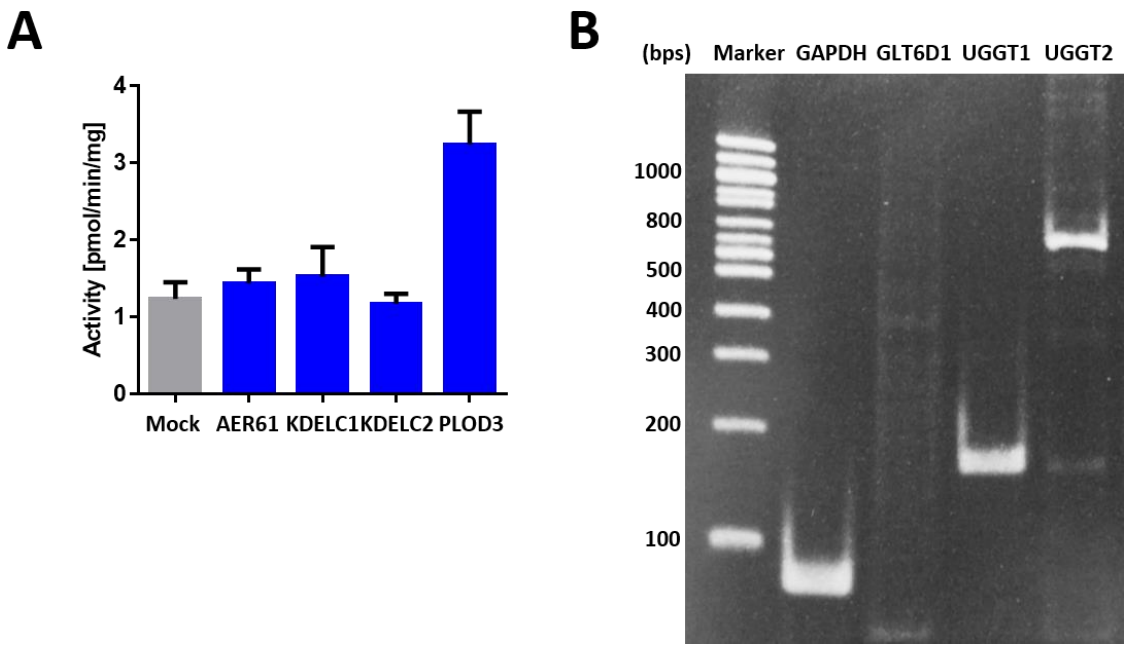


Figure 13| **ColGlcT activity for AER61, KDELC1, KDELC2, and PLOD3.** (A) PLOD3 shows some ColGlcT activity towards denatured type I collagen. (B) mRNA expression of GLT6D1, UGGT1, and UGGT2 in human fibroblast cells.

Another family found on the 34-list is the GT90 family which comprises three enzymes that are potential glucosyltransferases. One of them is the core glucosyltransferase protein O-β-glucosyltransferase Rumi or POGLUT. POGLUT has been described firstly in Drosophila and has now been shown to also have xylosyltransferase activity [7]. The other two enzymes have 55% identity in the amino acid sequence and are named KDEL (Lys-Asp-Glu-Leu) containing 1 and 2 (KDELC1 and KDELC2). No function has been assigned to both enzymes. We identified KDELC1 in the ColGlcT active fraction purified from chicken embryo extracts. We cloned both the KDELC1 and KDELC2 into the baculovirus expression system and tested for collagen glucosyltransferase activity. No glucosyltransferase activity could be measured towards collagen (Figure 13A). Additionally, for neither KDELC1 nor KDELC2 a homologues protein exists in *C. elegans* or *A. queenslandica* (Figure 15). Human KDELC1 shares the highest similarity in the *A. queenslandica* proteome with POGLUT by 33% homology.

GLT6D1 is an unknown member of the possibly α-glycosyltransferase group GT6. Besides the enzyme's unknown function, GLT6D1 has been found to possess glucosyltransferase activity in an

assay including the substrate gelatin (non-published report). Gelatin features the ColGlcT acceptor site Gal-Hyl contemplating a possible collagen glucosyltransferase activity. However, blasting GLT6D1 did not reveal conservation of the gene throughout the animal kingdom. No homologue was identified in either *C. elegans* or *A. queenslandica* proteomes. Beyond that, GLT6D1 mRNA is not expressed in human fibroblasts producing collagen indicating GLT6D1 is not the ColGlcT (Figure 13B).

We expanded the bioinformatics approach and performed extensive blast searches of all known and predicted glycosyltransferases from the proteomes of *Homo sapiens*, the nematode *Caenorhabditis elegans*, the sponge *Amphimedon queenslandica*, the fruitfly *Drosophila melanogaster,* the pea aphid *Acyrthosiphon pisum,* the monarch butterfly *Danaus plexippus,* the bdelloid rotifer *Adineta vaga* and the yeast *Saccharomyces cerevisiae* (Figure 15). The GTs were grouped according to the CAZy families. Only the GT-families containing one or more gene entries for *H. sapiens* were considered and GT-families that do not contain at least a single gene for each species analysed, were removed from the list. Except for *A. vaga* and *S. cerevisiae*, because both of them lack collagen but also have the processing enzymes of the N-glycosylation pathway. By comparing the resulting list with the list of the 34 unknown or uncharacterized human GTs, we could exclude the above described candidate GTs GLT8D1, GLT8D2 and KDELC1 since they are not conserved among the animal kingdom (Figure 15). On the other hand, we were looking more detailed at GTDC1, B3GALTL, and UGGT2. All of them were not well characterized and have homologues in *A. queenslandica*. For the Glycosyltransferase-like domain-containing protein 1 (GTDC1) we could not find a homologues protein in *C. elegans*. The closest similarity has been identified to the α-mannosyltransferase ALG2 which is also a GT4 enzyme. The GT31 enzyme beta-

Glycosyltransferase homologues among the animal kingdom

	Human	Murine	Bird	Fruit fly	Sawfly	Nematode	Platelminthes	Sponge	Rotifera	Yeast	Virus
	H.sapiens	**M.musculus**	**G.gallus**	**D.melanogaster**	**N.ribesii**	**C.elegans**	**S.japonicum**	**A.queenslandica**	**A.vaga**	**S.cerevisiae**	**A.mimivirus**
Collagen	Yes	Yes	Yes	Yes	Yes (silk)	Yes	Yes	Yes	No	No	Yes
O-glycan on Hyl	Glc, Gal	Glc, Gal	Glc, Gal	Glc, Gal	No	Gal	??	Glc, Gal	No	No	Glc
N-glycan (OST)	DDOST	Ddost	DDOST	OST48	-	T09A5.11	OST	OST48	-	WBP1(OSTB)	-
Protein UGGT	UGGT1, UGGT2	Uggt1, Uggt2	UGGT1, UGGT2	UGGT	-	UGGT-1, UGGT-2	UGGT, partial	UGGT1-like, UGGT-like	-	Kre5p	-
Protein PLOD3	PLOD3	Plod3	-	PLOD	-	Let-268	PLOD	PLOD3-like	-	-	L230
Protein COLGALT	GLT25D1, GLT25D2	Glt25d1, Glt25d2	GLT25D1, GLT25D2	CG31915	-	D2045.9	GLT25D2	GLT25D2-like	-	-	L230
Protein GTDC1	GTDC1	Gtdc1	GTDC1	GTDC1	-	-	-	GTDC1-like	-	-	
Protein GLT6D1	GLT6D1	Glt6d1	-	-	-	-	-	-	-	-	
Protein GLT8D2	GLT8D2	Glt8d2	GLT8D2	-	-	-	-	-	-	-	
Protein KDELC1	KDELC1	Kdelc1	KDELC1	-	-	-	-	-	Yes	-	
Protein POGLUT	POGLUT1	Poglut1	POGLUT1	Rumi	-	-	-	POGLUT-like	Yes	-	

The data presented is a collection from (Spiro 1967, Spiro 1971, Katzmann 1972, and uniprot.org search) and blast search for the indicated proteins against the respective organism from NCBI database.

Figure 14| **Homology of selected glycosyltransferases in the animal kingdom.** Blast searches of human glycosyltransferases against the indicated proteomes reveal that GLT6D1, GLT8D2, and KDELC1 are not so widespread among the animal kingdom than the known collagen modifying enzymes PLOD3 and ColGalT. POGLUT1 is characterized and described but appears in this list since it shares 33% homology to the unknown KDELC1. Besides the prominent abundance of the UGGT enzymes, GTDC1 appears to be the most conserved glycosyltransferase but has no homologue in the nematode and platelminthes.

GT	Amphimedon queenslandica	Homo sapiens	Caenorhabditis elegans	Drosophila melanogaster	Acyrthosiphom pisum	Danaus plexippus	Adineta vaga	Saccharomyces cerevisiae S288c	A. queenslandica glycosyltransferases	H. sapiens glycosyltransferases
1	4	35	78	37	75	47	49	3	Alg13, Alg14	Alg13, Alg14; ceramidGalT; Ugt (GlcA)
2	7	6	6	5	4	7	32	5	Alg5;DPM1;HAS3;B3GNT8 (B3GNTL1)	Alg5;DPM1;HAS3;B3GNT8 (B3GNTL1)
3	1	2	1	1	3	1	4	2	GYS1	GYS1,2
4	3	6	4	4	3	4	15	3	Alg2; PIGA; **GTDC1**	Alg2, Alg11; PIGA; **GTDC1**
7	6	15	5	5	8	8	43	0	CHSY3; B4GALNT3; B4GALT7	CHSY1; B4GALT
8	1	9	6	3	6	3	11	3	GYG1	GYG1, GYG2; LARGE1, LARGE2; GLT8D 1 – 4; XXYLT
10	4	8	5	5	7	11	34	0	FUT 6,7,**10,11**	FUT 3 – **11**
13	2	2	5	1	4	1	9	0	Mgat1; POMGNT1	Mgat1; POMGNT1
14	3	11	20	1	1	1	6	0	GCNT2; XYLT2	GCNT (core GlcNAc); XYLT1,2
16	0	1	1	1	3	6	4	0		Mgat2
21	1	1	3	1	1	1	4	0	CEGT	CEGT
22	3	4	3	4	5	6	9	4	Alg9, Alg12; PIGB	Alg9, Alg12; PIGB, PIGZ
23	0	1	1	1	10	1	6	0		FUT 8 (core α1-6)
24	8	2	2	1	1	2	2	1	UGGT **(UGGT2)**	UGGT1, **UGGT2**
25	4	3	1	1	1	1	1	0	GLT25D2	GLT25D 1 – 3
27	4	22	9	15	9	13	33	0	GALNT1, GALNT2, GALNT14	ppGALNT
31	12	26	30	23	9	12	51	0	CHSY1; B3GALT6; LFNG; **B3GALTL (b1-3 GlcT)**	GnT; GalT; Fringe; **B3GALTL (b1-3 GlcT)**
33	1	5	1	1	1	1	2	1	Alg1	Alg1; 4 fragments
35	1	3	1	1	2	1	10	1	PIGM	PIGM (muscle, brain, liver)
39	2	2	0	2	3	2	0	7	PomT1, PomT2	PomT1, PomT2
41	1	1	1	1	7	2	2	0	OGT	OGT
43	8	3	7	3	4	2	2	0	B3GAT2, B3GAT3	B3GAT 1 – 3
47	3	4	2	3	2	3	2	**0**	EXT1, EXT2, Extlike3	EXT1, EXT2 ; Extlike 1 – 3
57	2	2	2	2	0	2	4	2	Alg6, Alg8	Alg6, Alg8
58	2	1	1	1	0	1	3	1	Alg3	Alg3
59	1	1	1	1	0	1	2	1	Alg10	Alg10
61	2	2	1	1	1	2	3	0	AER61 (EOGT)	AER61 (EOGT); AGO61 (POMGNT2, GTDC2)
64	5	5	1	3	2	4	1	0	EXT1, EXT2, **Extlike3**	EXT1, EXT2 ; **Extlike 1 – 3**
65	1	1	1	1	1	1	6	0	PoFUT1	PoFUT1 (FUT 12)
66	4	2	1	2	4	2	4	1	STT3B	STT3A, STT3B
68	1	1	1	1	1	1	2	0	PoFUT2	PoFUT2 (FUT 13)
76	1	1	1	1	1	1	2	1	PIGV	PIGV (GPI-Man2)
90	1	2	0	2	2	3	4	0	POGLUT1 (RUMI)	POGLUT1 (RUMI); KDELC1

Figure 15| **Proteome analysis of CAZy glycosyltransferase families from sponge to human**. The analysis was performed by Bernard Henrissat from the CAZy glycogenomics team.

1,3-glucosyltransferase (B3GALTL) has been identified once in ColGlcT active fractions purified from chicken embryo homogenates and is found in *A. queenslandica* and *C. elegans* where it is named ZC250.2. We did not identify B3GALTL repeatedly and termed the enzyme unlikely to be the ColGlcT since it has been reported that B3GALTL glucosylates O-linked fucosylglycans on thrombospondin type 1 repeat domains [8].

SCREENINGS OF UNTYPED CONNECTIVE TISSUE DISORDER CASES FOR GLYCOSYLATION DEFECTS

Fibroblast cells from patients with unknown connective tissue disorders were screened for collagen glycosylation deficiencies. The patients showed characteristic clinical features affecting the skin and bone as observed in osteogenesis imperfecta and brittle bone disease patients. The collagen cross-links lysyl-pyridinoline (LP) and hydroxylysyl-pyridinoline (HP) ratio LP/HP were reduced in some of these patients indicating over modified collagen. Additional analysis performed at the Kinderspital Zürich revealed normal levels for LH and P4H activities. The finding of over modified collagens despite normal LH and P4H activities concluded a rate limiting step for modifying enzymes. Defects in the glycosylating enzymes could inhibit normal collagen triple helix formation thereby prolonging the hydroxylation event. Conceivably, defective collagen glycosylation could also hamper collagen secretion and proper processing. To identify whether collagen glycosylation is affected we tested the enzymatic activity by measuring the collagen glycan elongation using [^{14}C]-labeled Glc and Gal. Collagen glycosyltransferase assays were generally performed as described in *Schegg et al.* [9] using UDP-[^{14}C]Glc and UDP-[^{14}C]Gal (20 µCi/ml) (Perkin-Elmer Life Sciences) as donor and heat denatured collagen type I (250 µg, 10 min at 60 °C) as acceptor. All patient fibroblast cell lysates that were analyzed exhibit collagen glycosyltransferase activity (Figure 15 and 16). Some causes of disease were later identified as genetic defects of the P3H complex proteins CRTAP and P3H1.

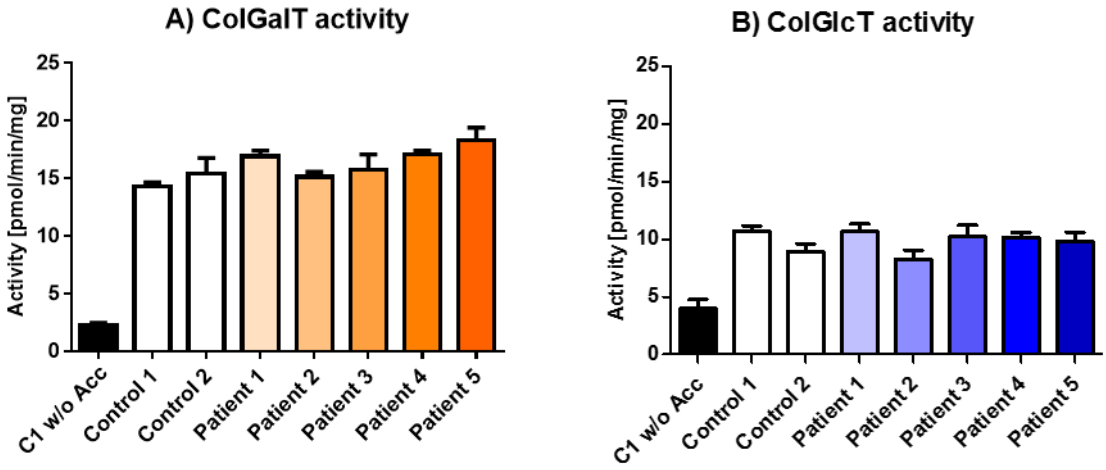


Figure 15| **Enzymatic activity of untyped EDS or OI cases.** Fibroblast cell lysates from 5 untyped EDS or OI patients were tested for ColGalT activity (A) and ColGlcT activity (B). Denatured bovine collagen type I was used as acceptor substrate in panel A and was galactosylated with GLT25D1 prior to serve as an acceptor for panel B assays. "w/o Acc" means the assay was performed without acceptor substrate. n = 3. Errorbars indicate mean ± SEM.

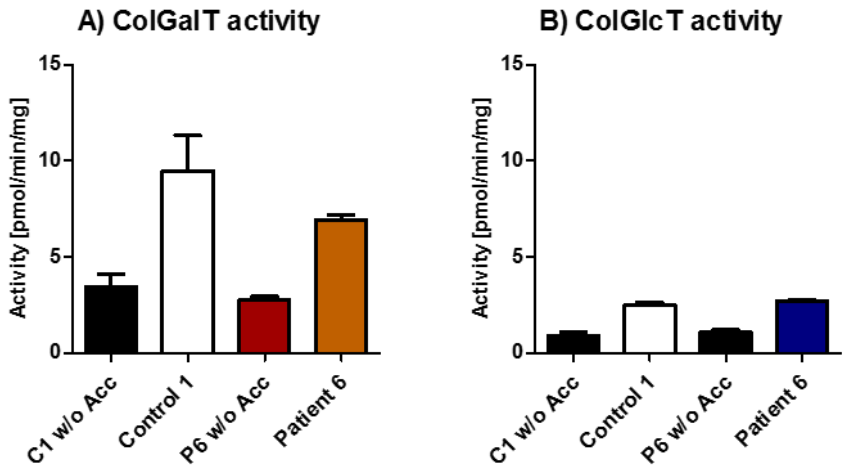


Figure 16| **Enzymatic activity of an untyped connective tissue disorder case.** Fibroblast cell lysates from an untyped connective tissue disorder patient was tested for ColGalT activity (A) and ColGlcT activity (B). In both assays denatured bovine collagen type I was used as acceptor substrate or when indicated without substrate (w/o Acc). n = 2. Errorbars indicate mean ± maximal deviation from mean.

Patient fibroblast cells were kindly provided by Prof. Dr. Matthias Baumgartner from the Kinderspital Zürich and by Dr. M. Rubio from the metabolic disease institute in Maastricht, Netherlands.

BIOTECHNOLOGICAL APPLICATION OF MIMIVIRAL COLLAGEN MODIFYING ENZYMES (PUBLICATION)

Recombinant expression of hydroxylated human collagen in *Escherichia coli*

Christoph Rutschmann[#], Stephan Baumann[#], Jürg Cabalzar, Kelvin B. Luther,

and Thierry Hennet[1]

Institute of Physiology, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

[#]both authors contributed equally

[1]To whom correspondence should be addressed: Thierry Hennet, Institute of Physiology, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland, Tel: +41 44 635 5080; Fax: +41 44 635 6814; E-mail: thennet@access.uzh.ch

ABSTRACT

Collagen is the most abundant protein in the human body and thereby a structural protein of considerable biotechnological interest. The complex maturation process of collagen, including essential post-translational modifications such as prolyl and lysyl hydroxylation, has precluded large-scale production of recombinant collagen featuring the biophysical properties of endogenous collagen. The characterization of new prolyl and lysyl hydroxylase genes encoded by the giant virus mimivirus reveals a method for production of hydroxylated collagen. The coexpression of a human collagen type III construct together with mimivirus prolyl- and lysyl hydroxylases in *Escherichia coli* yielded up to 90 mg of hydroxylated collagen per liter culture. The respective levels of prolyl and lysyl hydroxylation reaching 25% and 26% were similar to the hydroxylation levels of native human collagen type III. The distribution of hydroxyproline and hydroxylysine along recombinant collagen was also similar to that of native collagen as determined by mass spectrometric analysis of tryptic peptides. The triple helix signature of recombinant hydroxylated collagen was confirmed by circular dichroism, which also showed that hydroxylation increased the thermal stability of the recombinant collagen construct. Recombinant hydroxylated collagen produced in *Escherichia coli* supported the growth of human umbilical endothelial cells, underlining the biocompatibility of the recombinant protein as extracellular matrix. The high yield of recombinant protein expression and the extensive level of prolyl and lysyl hydroxylation achieved indicate that recombinant hydroxylated collagen can be produced at large scale for biomaterials engineering in the context of biomedical applications.

Keywords: Protein engineering, post translational modification, hydroxylysine, hydroxyproline, virus

INTRODUCTION

The structural and functional versatility of collagen in vertebrates makes it a coveted protein for tissue and biomaterials engineering. Yet, the considerable size of collagen polypeptides and the requirement for post-translational modifications have impeded the large-scale production of recombinant collagen featuring the biophysical properties of natural collagen. All types of collagen feature a triple helical conformation composed of repeats of the G-x-y motif, in which proline and lysine often occur at the x and y positions. During translation in the endoplasmic reticulum, selected proline and lysine residues are hydroxylated by dedicated hydroxylases, thereby yielding hydroxyproline (Hyp) and hydroxylysine (Hyl) (Myllyharju and Kivirikko 2004). The formation of Hyp is essential to stabilize the collagen triple helix and confer its thermal stability at body temperature (Shoulders and Raines 2009). Lysyl hydroxylation is involved in the formation of covalent intra- and inter-molecular crosslinks, contributing to condensation and fibril formation (Takaluoma et al. 2007). Hyl also serves as acceptor for the attachment of collagen-specific glycans (Schegg et al. 2009). Defects of lysyl hydroxylation lead to diseases such as Ehlers-Danlos type-VI (Hyland et al. 1992), Bruck syndrome (van der Slot et al. 2003), and skeletal dysplasia (Salo et al. 2008), demonstrating the biological importance of this post-translational modification.

The multimeric organization and limited stability of animal prolyl 4-hydroxylases and lysyl hydroxylases make them poor choices for the efficient production of recombinant collagen in conventional protein expression systems such as bacteria and yeasts, which lack endogenous prolyl and lysyl hydroxylases. Human collagen prolyl 4-hydroxylase has been expressed in *Escherichia coli* (Neubauer et al. 2005; Pinkas et al. 2011), although with limited activity towards short collagenous substrates. The coexpression of human prolyl 4-hydroxylase subunits and collagen constructs in the yeast *Pichia pastoris* has enabled the production of prolyl hydroxylated collagen up to 1.5 g per liter of culture (Nokelainen et al. 2001). Dual hydroxylation of proline and lysine has not yet been achieved in *Pichia pastoris*. Coexpression of human prolyl 4-hydroxylase subunits and the lysyl hydroxylase LH3 has been described in tobacco plants, in which recombinant human collagen type I was expressed at up to 200 mg per kg of fresh leaves (Stein et

al. 2009). The use of animal cells, such as Sf9 insect cells (Lamberg et al. 1996; Tomita et al. 1995) and HEK293 human cells (Fichard et al. 1997) that express prolyl and lysyl hydroxylase endogenously, yields recombinant collagen in the μg to mg range per liter of culture, thereby precluding large-scale applications of the recovered collagen product.

The description of several aquatic giant viruses belonging to *Phycodnaviridae* (Van Etten 2003) and *Mimiviridae* (Raoult et al. 2004) has shown that collagen-like genes are not restricted to metazoans and some prokaryotes. In addition to collagen genes, these viruses harbor genes encoding prolyl 4-hydroxylase (Eriksson et al. 1999) and lysyl hydroxylase enzymes (Luther et al. 2011). These viral hydroxylases are soluble and active when expressed in *E. coli*, thus opening new possibilities for the production of recombinant hydroxylated collagen in bacterial expression systems. So far, human collagen type II has been produced at amounts exceeding 10 g per liter (Guo et al. 2010), although without post-translational modifications. To circumvent this limitation, we now exploit bacterially active prolyl and lysyl hydroxylase enzymes from the giant virus mimivirus (Luther et al. 2011) to produce recombinant hydroxylated collagen at high yield in *E. coli*.

MATERIALS AND METHODS

Cloning of mimivirus hydroxylase expression vectors - The mimivirus lysyl hydroxylase L230 (Luther et al. 2011) and prolyl 4-hydroxylase L593 open reading frames were amplified by PCR from mimivirus genomic DNA using primers including *Xho*I and *Bam*HI sites. The primers were 5'-TGACCTCGAGATTAGTAGAACTTATGTAATT-3' and 5'-CAGGGATCCGTCCAATAAAGTGTAT-CAAC-3' for L230, 5'-TGACCTCGAGAAAACTGTGACTATCATTACAATA-3' and 5'-CAGGGA-TCCATTTTGTGTTAAAAAAATTTTAGG-3' for L593. The resulting amplicons were ligated as *Xho*I-*Bam*HI fragments into the *Xho*I-*Bam*HI linearized expression vector pET16b, yielding the pET16b-L230 and pET16b-L593 vectors. Expression vectors lacking His-tags were prepared by first amplifying the L230 and L593 genes using the primers 5'GTCGACG-AGCTCACCATGGGCATTAGTAGAAC-3' and 5'-GTAATGACATATGCGCAAGCCCAG-3' for L230, 5'-ATACCATGGTATTGTCAAAATCTTGTGTGT-3' and 5'-CAGGGATCCATTTTGTGTTAAAAA-AATTTTAGG-3' for L593. The corresponding amplicons were introduced into pET16b linearized with *Nco*I-*Nde*I for L230 and with *Nco*I-*Bam*HI for L593, yielding pET16b-noH-L230 and pET16b-noH-L593. The bicistronic vector pET16b-L593/L230 was prepared by inserting the expression cassette of the pET16b-L593 as a *Bgl*II-*Hind*III fragment into the *Bam*HI-*Hind*III-linearized pET16b-L230 vector. The bicistronic vector pET16b-noH-L593/L230 featuring L230 and L593 without His-tag was prepared in the same way.

Cloning of collagen expression vectors - A fragment of human collagen type III *COL3A1* cDNA encompassing 1206 bp and lacking propeptide-encoding regions was custom synthesized (GenScript, Piscataway, NJ, USA) using codons optimized for bacterial expression and including *Nco*I and *Bam*HI sites at 5'- and 3'-ends (Fig. 1). The pET28a expression vector was first digested with *Nco*I-*Bam*HI, which eliminates the His-tag at the N-terminal site. The resulting hCOL3 segment was inserted as a *Nco*I-*Bam*HI fragment into pET28a, yielding pET28a-hCOL3-His.

Protein expression in E. coli - The pET16b- and pET28a-based vectors were transformed into chemically competent *E. coli* BL21 (DE3) cells, which were plated on LB-agar plates containing

50 µg/ml kanamycin (Sigma-Aldrich) and 100 µg/ml ampicillin (Sigma–Aldrich) and incubated overnight at 37°C. Protein expression followed standard protocols (Tolia and Joshua-Tor 2006). Briefly, bacteria were grown in liquid cultures at 37 °C under agitation at 220 rpm until reaching an OD_{600} value of 0.6. Isopropyl β-D-1-thiogalactopyranoside (Sigma-Aldrich) was added to 1 mM to induce expression and the cultures were incubated for a further 3 h at 34 °C under agitation at 220 rpm.

Protein purification - Cells were pelleted at 4,000 x *g* at 4 °C for 30 min, resuspended in 3 ml of 20 mM sodium phosphate, pH 7.4, 100 mM NaCl per gram of *E. coli* wet weight and lysed with 250 µg/ml lysozyme, 4 mg/ml deoxycholic acid under rotation at 4 °C for 20 min. DNAse I (Fluka, Buchs, Switzerland) was added to 20 µg/ml and incubation proceeded at room temperature for 30 min. Cell lysates were clarified by centrifugation at 12,000 x *g* for 30 min at 4 °C and filtered through 0.22 µm membrane filters (Milipore). Imidazole was added to a concentration of 20 mM and His-tagged proteins were purified by affinity chromatography on a 1 ml HisTrap FF Ni Sepharose 6 column (GE - Healthcare) using an Äkta FPLC system (GE - Healthcare). Elution of His-tagged proteins was performed with 500 mM imidazole, 20 mM sodium phosphate, pH 7.4, 100 mM NaCl. His-tagged proteins were detected after transfer to nitrocellulose (Highbound ECL, GE-Healthcare) using the anti-polyHis HIS-1 monoclonal anti-body (Sigma-Aldrich).

Prolyl and lysyl hydroxylase activity assays - Prolyl and lysyl hydroxylase activities were measured as described previously (Luther et al. 2011). Briefly, 5 µg His-tag purified L230 lysyl hydroxylase or L593 prolyl hydroxylase were added to acceptor peptides at 0.5 mg/ml in 50 mM Tris-HCl, pH 7.4, 100 µM $FeSO_4$, 1 mM ascorbate, 100 µM DTT, 60 µM 2-oxoglutarate and 100 nCi of 2-oxo[^{14}C]glutarate (PerkinElmer Life Sciences) in a total volume of 100 µl and incubated at 37 °C for 45 min. Released [^{14}C]O_2 was captured in a filter paper soaked in NCS II

Tissue Solubilizer (GE Healthcare) suspended above the assay in a sealed 30 ml vial (VWR, Dietikon, Switzerland). Assays were stopped by addition of 100 µl ice-cold 1 M KH_2PO_4 and the filter papers were transferred to scintillation vials filled with 10 ml of IRGA-Safe Plus scintillation fluid (PerkinElmer Life Sciences). Radioactivity was measured in a Tri-Carb 2900TR scintillation counter (PerkinElmer Life Sciences).

Amino acid analysis - Purified collagens (10 µg) were hydrolyzed in 500 µl of 6 M HCl for 12 h at 105 °C. Hydrolysates were dried down under nitrogen, then washed twice with 500 µl of H_2O and dried down again. Samples were resuspended in 100 µl of H_2O and derivatized using 9-fluorenylmethoxycarbonyl chloride (FMOC) following the procedure of Bank *et al*. (Bank et al. 1996). Derivatized amino acid samples were analyzed by reverse phase HPLC as described in Schegg *et al.* (Schegg et al. 2009).

Mass spectrometry - Purified collagens (2 µg) were alkylated with iodoacetamide and digested with trypsin (Shevchenko et al. 2006). Briefly, after diluting the sample in 100 mM ammonium bicarbonate, 0.1 % (w/v) RapiGest (Waters, Saint-Quentin, France) and 5 mM dithiothreitol, the sample was heated for 30 min at 60 °C, cooled, and alkylated in 15 mM iodoacetamide for 30 min in the dark. Proteins were digested with trypsin overnight at 37 °C and acidified with trifluoroacetic acid to a final concentration of 0.5 % prior to desalting using a C18 ZipTip (Millipore). Tryptic digests were subjected to reverse phase LC-MS/MS analysis using a custom packed 150 mm x 0.075 mm Magic C18- AQ, 3 µm, 200 Å, column (Bischoff GmbH, Leonberg, Germany) and an Orbitrap Velos mass spectrometer (Thermo-scientific). Peptides were separated with an 80 min gradient of 3% to 97% of a buffer containing 99.8 % acetonitrile and 0.2 % formic acid. Spectra were recorded in the higher energy collisional dissociation mode acquiring 10 MS/MS spectra per MS scan with a minimal signal threshold of 2000 counts. Peptides were identified and assigned using Matrix Science Mascot version 2.4.1 and verified with the Scaffold version 4 software (Proteome Software, Inc.) using the X! Tandem search engine. Variable modifications included 16 Da on methionine, proline and lysine.

Circular Dichroism - Proteins were purified by gel filtration using a Superdex 200 10/300 GL Column (GE – Healthcare). Protein fractions were concentrated in a 10 kDa Spin-X^R UF 500 centrifugal concentrator (Corning) in PBS, pH 7.4, and kept at 4 °C at a concentration of 0.1 mg/ml prior to analysis. Human collagen type III was purchased from Sigma-Aldrich. Measurements were performed with a wavelength between 200 and 250 nm in a spectropolarimeter (J-810, Jasco) with a thermostated quartz cell of 1 mm length. Thermal stability was analyzed at 221.5 nm under heating at a rate of 0.5 °C/min from 4 °C to 70 °C.

Trypsin digestion of collagen - Recombinant hCOL3 (15 μg) in PBS pH 7.4 was digested with 15 ng trypsin (Roche) for 2 h at temperatures ranging from 10°C to 35°C. Digestions were stopped by addition of 2X Laemmli sample buffer and proteins were separated in 10% SDS-PAGE under reducing conditions.

Endothelial cell culture - Human umbilical vein endothelial cells (HUVEC) were cultured on 0.1% gelatin (Sigma-Aldrich), 0.1% recombinant hydroxylated hCOL3, 0.1% recombinant hCOL3 or 0.25% poly-D-lysine in ECM endothelial cell medium (ScienCell, Carlsbad, CA) at 37 °C in 5 % CO_2. For immunofluorescence, cells were seeded on glass cover slips at 1000 cells / cm^2, 13.3 μg coating matrix / cm^2 and cultured for 60 h. After washing twice with PBS, pH7.4, cells were fixed with 2 % paraformaldehyde for 10 min at room temperature, washed twice with 20 mM glycine in PBS and permeabilized with 1 mg/ml saponin. Cells were incubated with mouse anti-β-tubulin SAP.4G5 monoclonal antibody (Sigma-Aldrich) diluted 1:200 and labeled with rabbit anti-mouse IgG Alexa-488 (Life Technology) diluted 1:500 for 30 min. Nuclei were stained with DAPI (Biotium, Hayward, CA). Viability was assayed by methylthiazolyldiphenyl tetrazolium reduction using standard protocols (Mosmann 1983).

Sequence data - The L230 and L593 nucleotide sequences reported in this paper have the GenBank accession number NC_014649.1. The L230 and L593 protein sequences have the UniProtKB/Swiss-Prot accession numbers Q5UQC3 and Q5UP57, respectively. The nucleotide sequence of the human hCOL3 construct has the EMBL/EBI accession number HG779440.

RESULTS

The genome of the giant virus mimivirus contains seven collagen-like genes and open reading frames annotated as putative lysyl and prolyl hydroxylases (Raoult et al. 2004). We have previously demonstrated that the open reading frame L230 encodes a bifunctional collagen lysyl hydroxylase and glucosyltransferase enzyme (Luther et al. 2011). To confirm the activity of the putative proyl-4-hydroxylase encoded by the open reading frame L593, we expressed a His-tagged version of the protein in *E. coli*. The 669 bp open reading frame L593 yielded a 26 kDa protein, which could be enriched on Ni^{2+} beads (Fig. 2A). The prolyl hydroxylase activity of the purified L593 protein was assayed using acceptor peptides featuring proline in sequences derived from human collagen type I, type II, adiponectin and mannose-binding lectin. The L593 protein was active as prolyl hydroxylase on the artificial peptide sequence $(GPP)_7$ and on the peptides GDRGETGPAGPPGAPGAPGAP and GLRGLQGPPGKLGPPGNPGPS derived respectively from collagen type I and mannose binding lectin, each featuring the GPP motif (Fig. 2B). By contrast, prolyl hydroxylase activity was minimal on the peptides GPMGPSGPAGARGIQGPQGPR and GIPGHPGHNGAPGRDGRDGTP derived respectively from collagen type II and adiponectin, which lack the GPP motif (Fig. 2B). The L593 prolyl 4-hydroxylase was also active on the non-collagenous peptide $(SPAP)_5$ derived from proline-rich mimivirus proteins, thus indicating that L593 was not strictly specific towards G-x-y repeats (Fig. 2B).

To assess the ability of mimivirus L230 lysyl hydroxylase and L593 prolyl 4-hydroxylase to modify collagen fragments produced in *E. coli*, we coexpressed the two mimivirus hydroxylases together with a 38 kDa fragment of human COL3A1 collagen type III. To this end, the mimivirus L230 and L593 open reading frames were expressed bicistronically under kanamycin selection and the human hCOL3 fragment on a separate plasmid under ampicillin selection. The hCOL3 protein included 119 G-x-y repeats flanked by the N- and C-telopeptide sequences but lacking the N- and C-propeptide sequences (Fig. 1). The co-transformation of *E. coli* with the hydroxylase-containing plasmid and the human hCOL3 construct yielded expression of the three His-tagged target proteins at the expected molecular masses of 101 kDa, 37 kDa, and 26 kDa corresponding to L230,

hCOL3, and L593, respectively (Fig. 3A). As a next step, the L230 and L593 hydroxylases were expressed without His-tags to enable the single enrichment of the hCOL3 protein from *E. coli* cell lysates. The expression of hCOL3 alone or together with L230 and L593 hydroxylases showed that the collagen fragment reached similar expression levels. After purification by Ni^{2+} affinity chromatography, 90 mg of collagen protein per liter of bacterial culture were routinely obtained. Of note, the coexpression with L230 and L593 hydroxylases produced a smear above the expected hCOL3 band, suggestive of a larger protein size or a decreased migration in polyacrylamide gels (Fig. 3B). This smear may also reflect heterogeneity at the level of prolyl- and lysyl hydroxylation of the recombinant protein.

The level of prolyl- and lysyl hydroxylation of the hCOL3 protein achieved by co-expression with L230 and L593 hydroxylases was determined by amino acid analysis. Native human collagen type III and the recombinant His-tagged hCOL3 protein expressed with or without L230 and L593 hydroxylases were hydrolyzed under acidic conditions, and derivatized with FMOC. The separation of FMOC-labeled amino acids by HPLC analysis showed that 54% of proline residues and 47% of lysine residues were hydroxylated in native human collagen type III (Fig. 4A). By comparison, the levels of prolyl and lysyl hydroxylation reached respectively 25% and 26% in the human hCOL3 protein coexpressed with the L593 and L230 hydroxylases (Fig. 4B). In the absence of L593 and L230 hydroxylases no prolyl and no lysyl hydroxylation were observed in the recombinant hCOL3 protein (Fig. 4C). The efficient hydroxylation of recombinant hCOL3 indicates that substrates and co-factors required by the L593 and L230 hydroxylases are present in sufficient amounts in *E. coli* cultured in standard LB medium.

The distribution of Hyp and Hyl residues across the recombinant hCOL3 protein was investigated by mass spectrometry. The analysis of tryptic digested hCOL3 covered 92% of the sequence including 84 of 87 proline residues and all 12 lysine residues of the hCOL3 protein. The analysis of three different batches of recombinant hCOL3 protein revealed that between 66% and 83% of covered proline residues were hydroxylated (Table 1). For lysine residues, between 55% and 80% were detected as hydroxylated (Table 1). The assembly of tryptic peptides showed that

hydroxylation was evenly distributed across the hCOL3 protein (Fig. 5A). The mimivirus prolyl 4-hydroxylase did not appear to prefer proline residues at either the x or y position of the G-x-y motif. Several G-P-P motifs even included Hyp residues at both x and y positions. The recombinant hCOL3 protein included 12 lysine residues, of which 6 to 8 were hydroxylated (Fig. 5A). As observed for Hyp, the positions of Hyl residues within the G-x-y motif indicated that the mimivirus lysyl hydroxylase enzyme efficiently hydroxylated residues at both x and y positions. We compared the pattern of proline and lysine hydroxylation between native human collagen type III and the recombinant hCOL3 protein expressed in *E. coli*. The analysis revealed a similar distribution of hydroxylated amino acids across both polypeptide sequences (Fig. 5B). Overall, more Hyp residues were identified in the recombinant hCOL3 protein than in native collagen type III, although differences were minimal across the sequence regions surveyed. These sequences included only three lysine residues, only one of which was found to be hydroxylated in native collagen type III. By contrast, these three lysine residues were hydroxylated in the recombinant hCOL3 protein (Fig. 5B). Hyl residues on recombinant hCOL3 were not further modified, for instance, by glycosylation. We recently showed that the L230 protein is a bifunctional enzyme including both lysyl hydroxylase and Hyl glucosyltransferase domains (Luther et al. 2011). Whereas L230 efficiently converted lysine to Hyl, the enzyme failed to glycosylate the resulting Hyl residues on recombinant collagen, suggesting that the substrate UDP-Glc was not accessible in amounts sufficient to enable the L230-mediated glycosylation of recombinant collagen in *E. coli*.

The triple helical conformation and the thermal stability of the recombinant hCOL3 protein were investigated by circular dichroism. The ellipticity spectra obtained for non-hydroxylated and hydroxylated hCOL3 proteins showed the typical shape for triple helical collagen with a maximum peak around 221 nm and a negative peak below 200 nm (Fig. 6A). The changes in ellipticity at 221.5 nm during heating were monitored for non-hydroxylated and hydroxylated hCOL3 proteins to assess the thermal stability of both constructs. The triple helical conformation of the non-hydroxylated hCOL3 protein was unstable and showed an approximate 50% loss of ellipticity by 19.5°C (Fig. 6B). Since non-hydroxylated hCOL3 did not yield constant ellipticity values at low

temperatures, 19.5°C however cannot be defined as true Tm value. By contrast, the hydroxylated hCOL3 protein showed a 50% loss of ellipticity by 24.3°C, indicating that hydroxylation increased the thermal stability of the construct (Fig. 6B). We also compared the degree of triple helical conformation in non-hydroxylated and hydroxylated hCOL3 by digestion with trypsin, which cleaves denatured collagen but not triple helical collagen (Bruckner and Prockop 1981). Hydroxylated hCOL3 was resistant to trypsin up to 30°C whereas non-hydroxylated hCOL3 was mostly degraded by 30°C (Fig. 6C). Both forms of hCOL3 were completely degraded by 35°C, which confirmed their low thermal stability below 37°C.

The biocompatibility of hydroxylated and non-hydroxylated recombinant hCOL3 produced in *E. coli* was assessed by using the protein as a matrix supporting the growth of HUVEC. These cells prefer to grow on extracellular matrix proteins such as fibronectin and collagen (Smeets et al. 1992). The growth of HUVEC was compared between poly-D-lysine, bovine gelatin, recombinant non-hydroxylated hCOL3 and recombinant hydroxylated hCOL3 used as support. Cell morphology was examined by immunofluorescent staining of microtubules. When cultured on recombinant hydroxylated and non hydroxylated hCOL3, cell viability after 60 h culture reached respectively 64% and 49% of the viability observed when cells grew on 0.1% gelatin. As expected, viability was lowest when cells were cultured on poly-D-lysine (Fig. 7A). Cell morphology assessed by staining of microtubules showed that cells were evenly spread and tightly attached to the gelatin and hydroxylated hCOL3 matrices as indicated by the large number of processes (Fig. 7B). By contrast, the amount of rounded cells was elevated when non-hydroxylated hCOL3 was applied as matrix and few cells were visible when cultured on poly-D-lysine (Fig. 7B). The compatibility of hydroxylated hCOL3 as support for the growth of HUVEC demonstrated that the recombinant protein was suitable for biological applications such as matrix-assisted cell proliferation and adhesion.

DISCUSSION

The production of recombinant collagen requires post-translational modifications which are lacking in bacterial and yeast expression systems. In the present study we show that the prolyl hydroxylase L593 and the lysyl hydroxylase L230 from the giant virus mimivirus can be expressed as active enzymes in *E. coli* without any toxicity for the host cells. The coexpression of these two mimivirus hydroxylases with human collagen constructs enabled the efficient hydroxylation of proline and lysine residues across collagen requiring neither supplementation of co-factors, nor increased oxygen partial pressure.

To date, typical cost-effective and high-yield expression systems like yeasts and bacteria have not allowed the production of both prolyl and lysyl hydroxylated collagen because of the low activity of animal hydroxylases introduced in these hosts. The best results were obtained in the yeast *Pichia pastoris* expressing human prolyl 4-hydroxylase, which enabled 44% hydroxylation of proline residues on recombinant human collagen type III (Vuorela et al. 1997). However, efficient lysyl hydroxylation of collagen has not been achieved in *Pichia pastoris* so far. The degree of collagen hydroxylation is also a limiting factor for expression systems based on animal cells. Accordingly, the endogenous prolyl 4-hydroxylase and lysyl hydroxylase activities of insect cells did not yield efficient modification of recombinantly expressed collagen without co-transfection with human prolyl 4-hydroxylase subunits (Lamberg et al. 1996).

The expression of prolyl and lysyl hydroxylase enzymes derived from the giant virus mimivirus yielded degrees of hydroxylation for recombinantly expressed collagen close to those of native collagen type III. The roles of the prolyl hydroxylase L593 and lysyl hydroxylase L230 in mimivirus biology are unknown but the presence of seven collagen genes in the mimivirus genome (Raoult et al. 2004) suggests that L593 and L230 are involved in the hydroxylation of mimivirus collagen. Indeed, we have previously shown that the mimivirus collagen-like protein L71 is hydroxylated and glycosylated *in vitro* by the L230 enzyme (17). Structurally related proteins are also found in related giant viruses, such as megavirus (Arslan et al. 2011) and mou-mouvirus (Yoosuf et al. 2012), which also include collagen genes in their genome. Considering

their stability and activity when expressed in *E. coli*, proteins from giant viruses may represent a valuable source of enzymes for biotechnological applications, as shown here for the production of hydroxylated recombinant collagen.

The distribution of Hyp and Hyl residues on recombinant hCOL3 showed that mimivirus hydroxylases were able to hydroxylate proline and lysine in various sequence contexts. The pattern of prolyl hydroxylation showed that proline at either position x or y of the motif G-x-y could be efficiently hydroxylated. Studies performed on synthetic peptides containing Hyp at positions x or y or both demonstrated that Hyp at position y strongly increases thermal stability (Jiravanichanun et al. 2006) whereas Hyp at position x destabilizes the triple helical conformation in Gly-Hyp-y repeats (Inouye et al. 1982). The presence of Hyp at both positions x and y by contrast, further stabilized the triple helical conformation of the peptides (Berisio et al. 2004). The detection of several Hyp residues at position x on various types of collagen makes it difficult to predict the positional impact of Hyp on the thermal stability of more complex polypeptides (Bann and Bachinger 2000; Buechter et al. 2003; Song and Mechref 2013). Although early work demonstrated that Hyp occurs exclusively at position y (Fietzek and Rauterberg 1975), recent studies showed that Hyp also occurs at position x in fibrillar collagens (Song and Mechref 2013; Weis et al. 2010).

In animal cells the C-propeptide domain of collagen is important for the initiation of triple helix formation in the endoplasmic reticulum and contributes to the solubility of the molecules along the secretory pathway (Boudko et al. 2012). The addition of trimerization domains to short collagen constructs, such as the bacteriophage T4 foldon domain at the C-terminus of a $[GPP]_{10}$ sequence, was reported to dramatically increase the thermal stability of the collagen construct (Boudko et al. 2002). Although advantageous in accelerating triple helix formation, the addition of propeptides in recombinant collagen constructs expressed in *E. coli* or *Pichia pastoris* later requires their removal for formation of fibrillar structures. This procedure is usually performed by pepsin digestion, which leaves the triple helical domain intact but also removes the short telopeptides sequences required for the registration of collagen molecules in order to form fibrils

(Capaldi and Chapman 1982). Therefore, we chose to produce an hCOL3 construct devoid of propeptides but including the telopeptides necessary for fibrillogenesis. The absence of propeptides did not affect the solubility of the recombinant hCOL3 protein and simplified down-stream processing by avoiding protease digestion and removal from purified collagen.

The simplicity of this mimivirus hydroxylase expression system enables the efficient post-translational hydroxylation of proteins containing collagen domains. In addition to the family of true collagens, several collagenous proteins like adiponectin, mannose-binding lectin and the surfactant proteins A and D can now be produced as hydroxylated proteins in *E. coli*. The high yield of bacterial expression combined with a high degree of prolyl and lysyl hydroxylation provides the framework for the large-scale production of recombinant collagens for human applications, in which animal collagens represent significant risks for allergic reactions and zoonotic disease transmission.

ACKNOWLEDGMENTS

We are grateful to Dr. Peter Gehrig and Dr. Jonas Gossmann at the Functional Genomics Center Zurich for their support with mass spectrometric analyses and Marek Whitehead for endotoxin determination. This work was supported by the University of Zürich and by the Swiss National Foundation grant 310030-129633 to TH and by the Research Credit of the University of Zurich to SB.

COMPETING FINANCIAL INTERESTS

The University of Zürich has filed a patent on the application of the mimivirus hydroxylases for biotechnology purposes.

REFERENCES

Arslan D, Legendre M, Seltzer V, Abergel C, Claverie JM (2011) Distant Mimivirus relative with a larger genome highlights the fundamental features of *Megaviridae*. Proc Natl Acad Sci U S A 108(42):17486-91

Bank RA, Jansen EJ, Beekman B, te Koppele JM (1996) Amino acid analysis by reverse-phase high-performance liquid chromatography: improved derivatization and detection conditions with 9-fluorenylmethyl chloroformate. Anal Biochem 240(2):167-76

Bann JG, Bachinger HP (2000) Glycosylation/Hydroxylation-induced stabilization of the collagen triple helix. 4-trans-hydroxyproline in the Xaa position can stabilize the triple helix. J Biol Chem 275(32):24466-9

Berisio R, Granata V, Vitagliano L, Zagari A (2004) Imino acids and collagen triple helix stability: characterization of collagen-like polypeptides containing Hyp-Hyp-Gly sequence repeats. J Am Chem Soc 126(37):11402-3

Boudko S, Frank S, Kammerer RA, Stetefeld J, Schulthess T, Landwehr R, Lustig A, Bachinger HP, Engel J (2002) Nucleation and propagation of the collagen triple helix in single-chain and trimerized peptides: transition from third to first order kinetics. J Mol Biol 317(3):459-70

Boudko SP, Engel J, Bachinger HP (2012) The crucial role of trimerization domains in collagen folding. Int J Biochem Cell Biol 44(1):21-32

Bruckner P, Prockop DJ (1981) Proteolytic enzymes as probes for the triple-helical conformation of procollagen. Anal Biochem 110(2):360-8

Buechter DD, Paolella DN, Leslie BS, Brown MS, Mehos KA, Gruskin EA (2003) Co-translational incorporation of trans-4-hydroxyproline into recombinant proteins in bacteria. J Biol Chem 278(1):645-50

Capaldi MJ, Chapman JA (1982) The C-terminal extrahelical peptide of type I collagen and its role in fibrillogenesis in vitro. Biopolymers 21(11):2291-313

Eriksson M, Myllyharju J, Tu H, Hellman M, Kivirikko KI (1999) Evidence for 4-hydroxyproline in viral proteins. Characterization of a viral prolyl 4-hydroxylase and its peptide substrates. J Biol Chem 274(32):22131-4

Fichard A, Tillet E, Delacoux F, Garrone R, Ruggiero F (1997) Human recombinant alpha1(V) collagen chain. Homotrimeric assembly and subsequent processing. J Biol Chem 272(48):30083-7

Fietzek PP, Rauterberg J (1975) Cyanogen bromide peptides of type III collagen: first sequence analysis demonstrates homology with type I collagen. FEBS Lett 49(3):365-8

Guo J, Luo Y, Fan D, Yang B, Gao P, Ma X, Zhu C (2010) Medium optimization based on the metabolic-flux spectrum of recombinant *Escherichia coli* for high expression of human-like collagen II. Biotechnol Appl Biochem 57(2):55-62

Hyland J, Ala-Kokko L, Royce P, Steinmann B, Kivirikko KI, Myllyla R (1992) A homozygous stop codon in the lysyl hydroxylase gene in two siblings with Ehlers-Danlos syndrome type VI. Nat Genet 2(3):228-31

Inouye K, Kobayashi Y, Kyogoku Y, Kishida Y, Sakakibara S, Prockop DJ (1982) Synthesis and physical properties of (hydroxyproline-proline-glycine)10: hydroxyproline in the X-position decreases the melting temperature of the collagen triple helix. Arch Biochem Biophys 219(1):198-203

Jiravanichanun N, Nishino N, Okuyama K (2006) Conformation of alloHyp in the Y position in the host-guest peptide with the pro-pro-gly sequence: implication of the destabilization of (Pro-alloHyp-Gly)10. Biopolymers 81(3):225-33

Lamberg A, Helaakoski T, Myllyharju J, Peltonen S, Notbohm H, Pihlajaniemi T, Kivirikko KI (1996) Characterization of human type III collagen expressed in a baculovirus system. Production of a protein with a stable triple helix requires coexpression with the two types of recombinant prolyl 4-hydroxylase subunit. J Biol Chem 271(20):11988-95

Luther KB, Hulsmeier AJ, Schegg B, Deuber SA, Raoult D, Hennet T (2011) Mimivirus collagen is modified by bifunctional lysyl hydroxylase and glycosyltransferase enzyme. J Biol Chem 286(51):43701-9

Mosmann TR (1983) Rapid colorimetric assay for cellular growth and survival. Application to proliferation and cytotoxicity assays. J Immunol Meth 65:55-65

Myllyharju J, Kivirikko KI (2004) Collagens, modifying enzymes and their mutations in humans, flies and worms. Trends Genet 20(1):33-43.

Neubauer A, Neubauer P, Myllyharju J (2005) High-level production of human collagen prolyl 4-hydroxylase in *Escherichia coli*. Matrix Biol 24(1):59-68

Nokelainen M, Tu H, Vuorela A, Notbohm H, Kivirikko KI, Myllyharju J (2001) High-level production of human type I collagen in the yeast *Pichia pastoris*. Yeast 18(9):797-806

Pinkas DM, Ding S, Raines RT, Barron AE (2011) Tunable, post-translational hydroxylation of collagen Domains in Escherichia coli. ACS Chem Biol 6(4):320-4

Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H, La Scola B, Suzan M, Claverie JM (2004) The 1.2-megabase genome sequence of Mimivirus. Science 306(5700):1344-50

Salo AM, Cox H, Farndon P, Moss C, Grindulis H, Risteli M, Robins SP, Myllyla R (2008) A connective tissue disorder caused by mutations of the lysyl hydroxylase 3 gene. Am J Hum Genet 83(4):495-503

Schegg B, Hülsmeier AJ, Rutschmann C, Maag C, Hennet T (2009) Core glycosylation of collagen is initiated by two beta(1-O)galactosyltransferases. Mol Cell Biol 29(4):943-952

Shevchenko A, Tomas H, Havlis J, Olsen JV, Mann M (2006) In-gel digestion for mass spectrometric characterization of proteins and proteomes. Nat Protoc 1(6):2856-60

Shoulders MD, Raines RT (2009) Collagen structure and stability. Annu Rev Biochem 78:929-58

Smeets EF, von Asmuth EJ, van der Linden CJ, Leeuwenberg JF, Buurman WA (1992) A comparison of substrates for human umbilical vein endothelial cell culture. Biotech Histochem 67(4):241-50

Song E, Mechref Y (2013) LC-MS/MS Identification of the O-Glycosylation and Hydroxylation of Amino Acid Residues of Collagen alpha-1 (II) chain from Bovine Cartilage. J Proteome Res 12(8):3599-609

Stein H, Wilensky M, Tsafrir Y, Rosenthal M, Amir R, Avraham T, Ofir K, Dgany O, Yayon A, Shoseyov O (2009) Production of bioactive, post-translationally modified, heterotrimeric, human recombinant type-I collagen in transgenic tobacco. Biomacromolecules 10(9):2640-5

Takaluoma K, Hyry M, Lantto J, Sormunen R, Bank RA, Kivirikko KI, Myllyharju J, Soininen R (2007) Tissue-specific changes in the hydroxylysine content and cross-links of collagens and alterations in fibril morphology in lysyl hydroxylase 1 knock-out mice. J Biol Chem 282(9):6588-96

Tolia NH, Joshua-Tor L (2006) Strategies for protein coexpression in *Escherichia coli*. Nat Methods 3(1):55-64

Tomita M, Ohkura N, Ito M, Kato T, Royce PM, Kitajima T (1995) Biosynthesis of recombinant human pro-alpha 1(III) chains in a baculovirus expression system: production of disulphide-bonded and non-disulphide-bonded species containing full-length triple helices. Biochem J 312 (Pt 3):847-53

van der Slot AJ, Zuurmond AM, Bardoel AF, Wijmenga C, Pruijs HE, Sillence DO, Brinckmann J, Abraham DJ, Black CM, Verzijl N, DeGroot J, Hanemaaijer R, TeKoppele JM, Huizinga TW, Bank RA (2003) Identification of PLOD2 as telopeptide lysyl hydroxylase, an important enzyme in fibrosis. J Biol Chem 278(42):40967-72

Van Etten JL (2003) Unusual life style of giant chlorella viruses. Annu Rev Genet 37:153-95

Vuorela A, Myllyharju J, Nissi R, Pihlajaniemi T, Kivirikko KI (1997) Assembly of human prolyl 4-hydroxylase and type III collagen in the yeast *Pichia pastoris*: formation of a stable enzyme tetramer requires coexpression with collagen and assembly of a stable collagen requires coexpression with prolyl 4-hydroxylase. EMBO J 16(22):6702-12

Weis MA, Hudson DM, Kim L, Scott M, Wu JJ, Eyre DR (2010) Location of 3-hydroxyproline residues in collagen types I, II, III, and V/XI implies a role in fibril supramolecular assembly. J Biol Chem 285(4):2580-90

Yoosuf N, Yutin N, Colson P, Shabalina SA, Pagnier I, Robert C, Azza S, Klose T, Wong J, Rossmann MG, La Scola B, Raoult D, Koonin EV (2012) Related giant viruses in distant locations and different habitats: *Acanthamoeba polyphaga* moumouvirus represents a third lineage of the *Mimiviridae* that is close to the megavirus lineage. Genome Biol Evol 4(12):1324-30

FIGURE LEGENDS

FIGURE 1. DNA and protein sequence of synthetic human hCOL3 construct. The top panel shows the codon optimized DNA sequence of the truncated human hCOL3 cDNA construct flanked by a 5' *Nco*I site and 3' *Bam*HI site (underlined). The ATG and TGA stop codon are bold and shaded. The sequence encoding the His-tag preceding the stop codon is dash-underlined. The bottom panel shows the amino acid sequence of the truncated human hCOL3 protein. The Gly-x-y collagen domain is shaded.

FIGURE 2. Bacterial expression and characterization of mimivirus L593. **A**, SDS-PAGE of mimivirus L593 expressed in *E. coli* shown as cell lysate (L) and after Ni^{2+}-affinity purification (P), either after staining with Coomassie blue R-250 or after Western blotting with anti-His_6 antibody. **B**, Prolyl hydroxylase activity of purified mimivirus L593 assayed on the peptide acceptors $[SPAP]_5$ (1), $[GPP]_7$ (2), GDRGETGPAGPPGAPGAPGAP from human collagen type I (3), GPMGPSGPAGARGIQGPQGPR from human collagen type II (4), GLRGLQGPPGKLGPPGNPGPS from human mannose-binding lectin (5), GIPGHPGHNGAPGRDGRDGTP from human adiponectin (6). Open bars show prolyl hydroxylase activity measured without peptide acceptor and black bars with peptide acceptors. Stars above bars indicate statistically significant activity using two-tailed paired t-test ($p<0.01$).

FIGURE 3. Coexpression of His-tagged mimivirus hydroxylases L593 and L230 with His-tagged human hCOL3 fragment. **A**, SDS-PAGE of mimivirus L593, mimivirus L230, and human hCOL3 construct expressed in *E. coli* shown as cell lysate (L) and after Ni^{2+}-affinity purification (P), either after staining with Coomassie blue R-250 or after Western blotting with anti-His_6 antibody. **B**, SDS-PAGE of His-tagged human hCOL3 construct expressed alone (-) or with L593 and L230

hydroxylases (+), shown after staining with Coomassie blue R-250 or after Western blotting with anti-His_6 antibody.

FIGURE 4. Amino acid analysis of native and recombinant human hCOL3. Purified hCOL3 proteins were acid hydrolyzed and the resulting amino acids labeled with FMOC, and separated by HPLC. The positions of amino acids are indicated by the single letter code. The positions of hydroxyproline (Hyp) and hydroxylysine (Hyl) are marked by arrows. **A**, native human COL3A1; **B**, recombinant hCOL3 construct coexpressed with mimivirus L593 and L230 hydroxylases; **C**, recombinant hCOL3 construct expressed alone.

FIGURE 5. Distribution of Hyp and Hyl on recombinant human hCOL3. **A**, The occurrence of hydroxylated residues was determined by mass spectrometric analysis of tryptic digests from recombinant human hCOL3 protein. Grayed sequences represent portions of the sequences not covered in the analysis. Proline (P) and lysine (K) residues identified as hydroxylated are shaded. **B**, Comparison of Hyp and Hyl distribution on stretches of native human COL3A1 (nat) and recombinant human hCOL3 (rec) produced in *E. coli*. Proline (P) and lysine (K) residues identified as hydroxylated are shaded.

FIGURE 6. Circular dichroism of recombinant human hCOL3. **A**, Samples of purified hydroxylated (left panel) and non-hydroxylated recombinant hCOL3 protein (right panel) protein at 0.1 mg/ml were scanned between 200 and 250 nm in a spectropolarimeter. **B**, Thermal transitions of hydroxylated (left panel) and non-hydroxylated recombinant hCOL3 protein (right panel) in PBS, pH7.4 measured at 221.5 nm under a heating rate of 0.5 °C/min from 4 °C to 70 °C. The T_m values were determined at the midpoints of the sigmoid curves. **C**, Trypsin digestion of hydroxylated (left

panel) and non-hydroxylated recombinant hCOL3 protein (right panel); the arrowhead at the right shows the position of the hCOL3 protein.

FIGURE 7. Growth of HUVEC on recombinant human hCOL3 matrix. **A**, The viability of HUVEC seeded at 1000 cells per cm^2 was determined by reduction of methylthiazolyldiphenyl tetrazolium to formazan after 60 h incubation at 37 °C on the matrices: 0.1% gelatin, 0.1% hydroxylated recombinant human hCOL3 (hCOL3-OH), 0.1% non-hydroxylated recombinant human hCOL3 (hCOL3), and 0.25% poly-D-lysine (PDL). Cell viability is expressed relatively to the values obtained for the positive control, 0.1% gelatin. Data show the average and standard error of the mean of three experiments. **B**, Immunofluorescence microtubule (green) and DNA (blue) staining of HUVEC grown on 0.1% gelatin, 0.1% hydroxylated recombinant human hCOL3 (hCOL3-OH), 0.1% non-hydroxylated recombinant human hCOL3 (hCOL3), and 0.25% poly-D-lysine (PDL).

Table 1. Hydroxylation efficiency of recombinant hCOL3 protein. Tryptic digests were analyzed for hydroxylation of proline (Pro) and lysine (Lys) by mass spectrometry.

		Coverage			Hyp/Pro			Hyl/Lys
	AA[a]	[%]	Pro	Hyp	[%]	Lys	Hyl	[%]
hCOL3	401		87			12		
Batch 1	346	86	80	56	70	11	6	55
Batch 2	343	86	80	53	66	11	6	55
Batch 3	291	73	63	52	83	10	8	80
Combination[b]	368	92	84	59	70	12	9	75

[a] covered amino acid length

[b] combined assembly of tryptic peptides from batches 1 to 3

FIGURE 1

```
   1  CCATGGATGT ATGATTCGTA TGATGTCAAG TCGGGTGTGG CAGTGGGTGG TCTGGCAGGC
  61  TATCCGGGTC CGGCAGGTCC GCCGGGTCCG CCGGGTCCGC CGGGTACCTC TGGTCATCCG
 121  GGTAGCCCGG GCTCTCCGGG TTATCAGGGT CCGCCGGGTG AACCGGGCCA AGCGGGTCCG
 181  AGCGGTCCGC CGGGTCCGCC GGGCGCTATT GGTCCGAGTG GCCCGGCGGG TAAAGATGGC
 241  GAATCCGGTC GTCCGGGTCG TCCGGGTGAA CGCGGCCTGC CGGGTCCGCC GGGTATTAAA
 301  GGTCCGGCAG GCATCCCGGG TTTTCCGGGT ATGAAGGGTC ACCGCGGCTT CGACGGTCGT
 361  AACGGCGAAA AAGGTGAAAC CGGTGCCCCG GGTCTGAAGG GTGAAAACGG TCTGCCGGGT
 421  GAAAATGGTG CTCCGGGTCC GATGGGTCCG CGTGGCGCGC CGGGTGAACG TGGTCGTCCG
 481  GGTCTGCCGG GTGCCGCAGG TGCCCGCGGC AACGATGGTG CACGTGGCAG TGACGGTCAG
 541  CCGGGTCCGC CGGGTCCGCC GGGGACCGCT GGTTTTCCGG GCTCCCCGGG TGCAAAAGGC
 601  GAAGTGGGTC CGGCAGGCAG TCCGGGTTCC AATGGTGCAC CGGGTCAGCG CGGCGAACCG
 661  GGTCCGCAAG GCCATGCCGG TCCGCCGGGC CCGGTTGGTC CGGCAGGCAA GAGCGGTGAT
 721  CGTGGCGAAT CTGGTCCGGC CGGTCCGGCT GGTGCCCCGG GTCCGGCCGG TAGTCGCGGC
 781  GCACCGGGTC CGCAAGGCCC GCGTGGTGAC AAAGGCGAAA CCGGTGAACG CGGCGCAGCT
 841  GGTATTAAGG GCCACCGTGG TTTCCCGGGC AATCCGGGTG CACCGGGCAG CCCGGGTCCG
 901  GCTGGCCAGC AGGGTGCCAT TGGCTCTCCG GGCCCGGCCG GTCCGCGTGG TCCGGTTGGT
 961  CCGTCAGGTC CGCCGGGTAA AGATGGCACG TCGGGTCATC CGGGTCCGAT TGGTCCGCCG
1021  GGTCCGCGTG GTAATCGCGG TGAACGTGGC TCAGAAGGTT CGCCGGGTCA CCCGGGCCAA
1081  CCTGGTCCGC CGGGTCCGCC GGGTGCTCCG GGTCCGTGCT GTGGCGGTGT TGGCGCGGCC
1141  GCAATCGCGG GCATCGGCGG CGAAAAGGCG GGCGGCTTTG CTCCGTATTA TCATCATCAC
1201  CATCACCATT GAGGATCC
```

```
   1  MYDSYDVKSG VAVGGLAGYP GPAGPPGPPG PPGTSGHPGS PGSPGYQGPP GEPGQAGPSG
  61  PPGPPGAIGP SGPAGKDGES GRPGRPGERG LPGPPGIKGP AGIPGFPGMK GHRGFDGRNG
 121  EKGETGAPGL KGENGLPGEN GAPGPMGPRG APGERGRPGL PGAAGARGND GARGSDGQPG
 181  PPGPPGTAGF PGSPGAKGEV GPAGSPGSNG APGQRGEPGP QGHAGPPGPV GPAGKSGDRG
 241  ESGPAGPAGA PGPAGSRGAP GPQGPRGDKG ETGERGAAGI KGHRGFPGNP GAPGSPGPAG
 301  QQGAIGSPGP AGPRGPVGPS GPPGKDGTSG HPGPIGPPGP RGNRGERGSE GSPGHPGQPG
 361  PPGPPGAPGP CCGGVGAAAI AGIGGEKAGG FAPYYHHHHH H
```

FIGURE 2

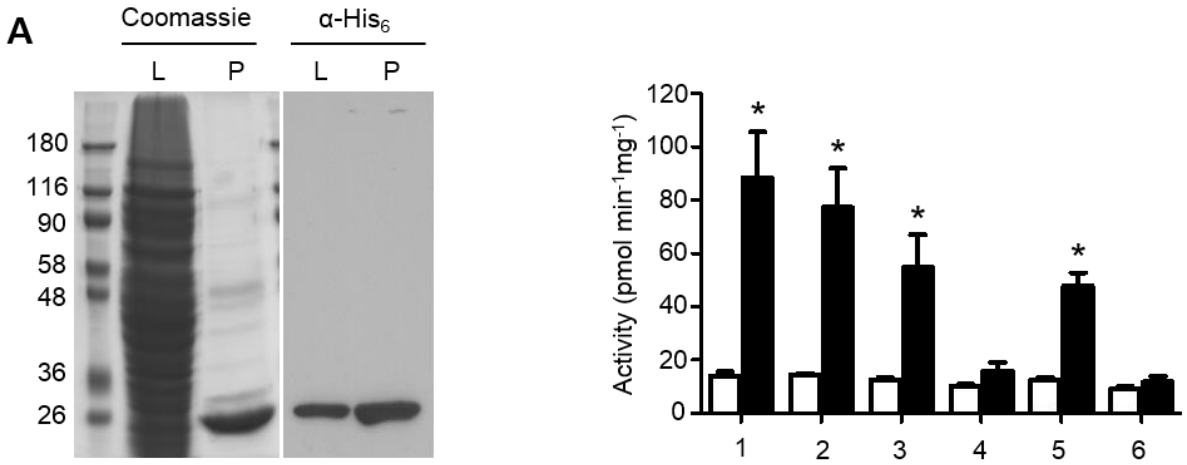


FIGURE 3

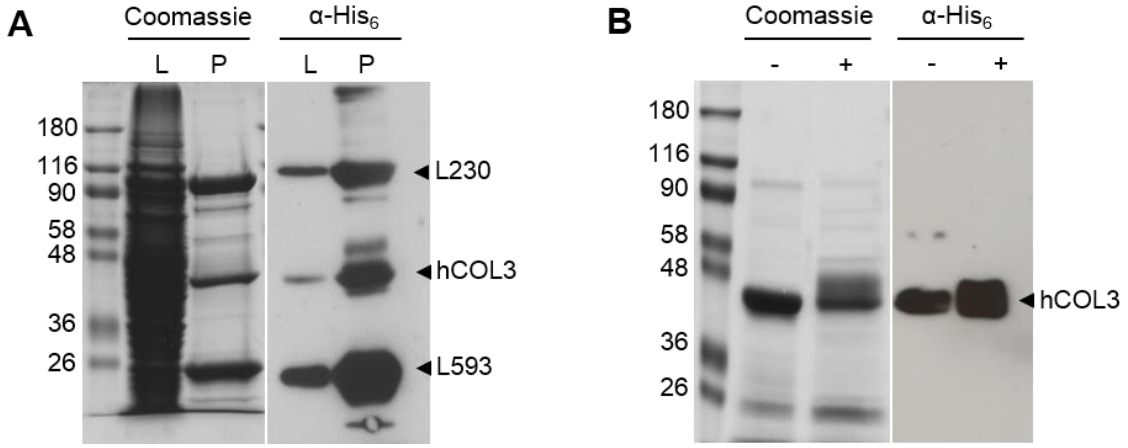


FIGURE 4

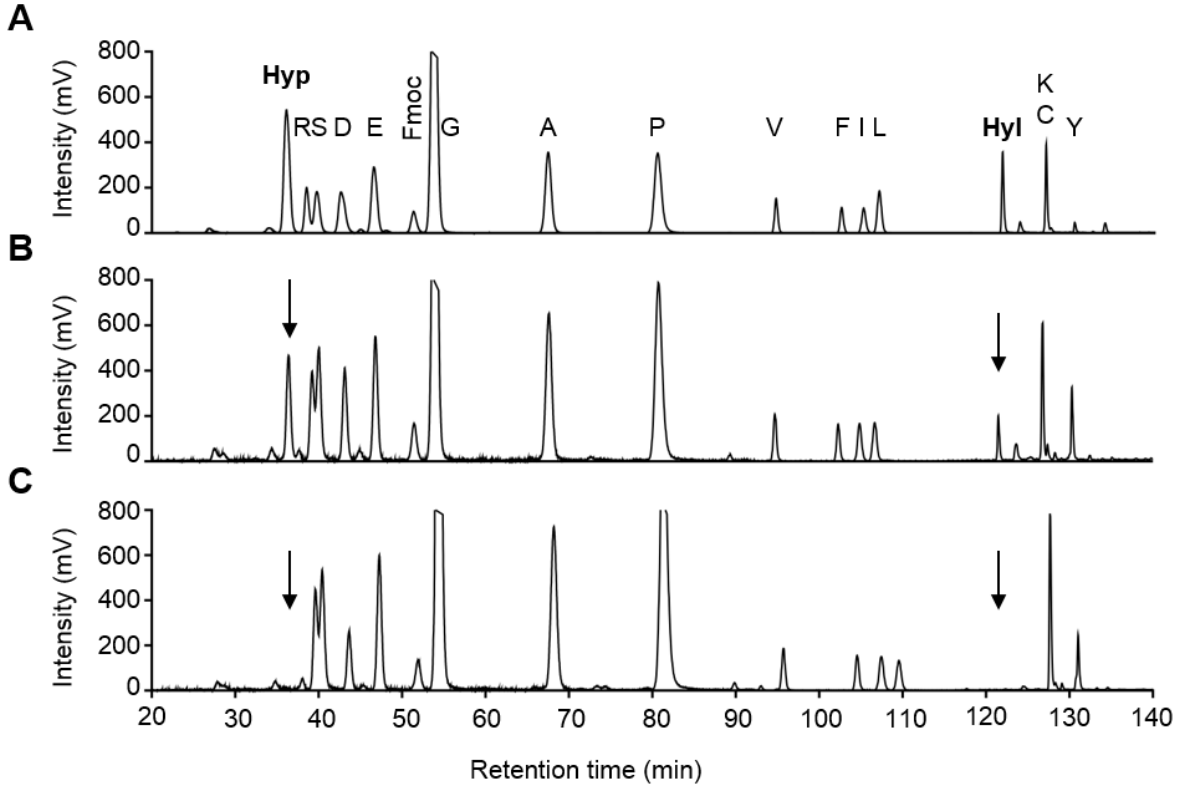


FIGURE 5

A

```
   1 MYDSYDVKSGVAVGGLAGYPGPAGPPGPPGPPGTSGHPGSPGSPGYQGPP
  51 GEPGQAGPSGPPGPPGAIGPSGPAGKDGESGRPGRPGERGLPGPPGIKGP
 101 AGIPGFPGMKGHRGFDGRNGEKGETGAPGLKGENGLPGENGAPGPMGPRG
 151 APGERGRPGLPGAAGARGNDGARGSDGQPGPPGPPGTAGFPGSPGAKGEV
 201 GPAGSPGSNGAPGQRGEPGPQGHAGPPGPVGPAGKSGDRGESGPAGPAGA
 251 PGPAGSRGAPGPQGPRGDKGETGERGAAGIKGHRGFPGNPGAPGSPGPAG
 301 QQGAIGSPGPAGPRGPVGPSGPPGKDGTSGHPGPIGPPGPRGNRGERGSE
 351 GSPGHPGQPGPPGPPGAPGPCCGGVGAAAIAGIGGEKAGGFAPYYHHHHH
 401 H
```

B

```
nat  308 GRPGLPGAAGARGNDGARGSDGQPGPPGPPGTAGFPGSPGAKGEVGPAGS
rec  156 GRPGLPGAAGARGNDGARGSDGQPGPPGPPGTAGFPGSPGAKGEVGPAGS

nat      PGSNGAPGQRGEPGPQGHA  376
rec      PGSNGAPGQRGEPGPQGHA  224


nat 1039 GPPGPVGPAGKSGDRGESGPAGPAGAPGPAGSRGAPGPQGPR 1080
rec  225 GPPGPVGPAGKSGDRGESGPAGPAGAPGPAGSRGAPGPQGPR  266


nat 1109 GFPGNPGAPGSPGPAGQQGAIGSPGPAGPRGPVGPSGPPGKDGTSGHPGP
rec  285 GFPGNPGAPGSPGPAGQQGAIGSPGPAGPRGPVGPSGPPGKDGTSGHPGP

nat      IGPPGPR 1165
rec      IGPPGPR  341
```

FIGURE 6

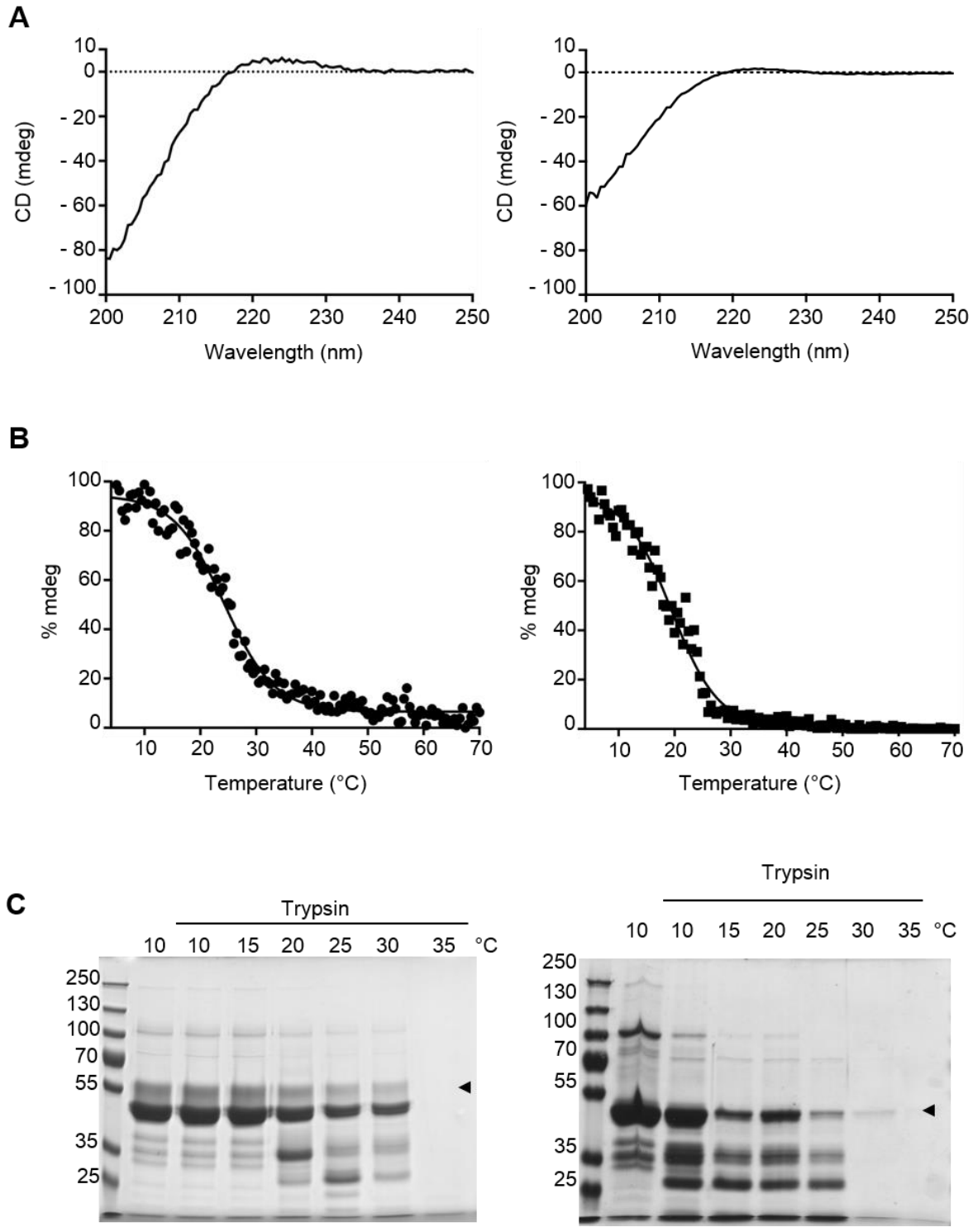
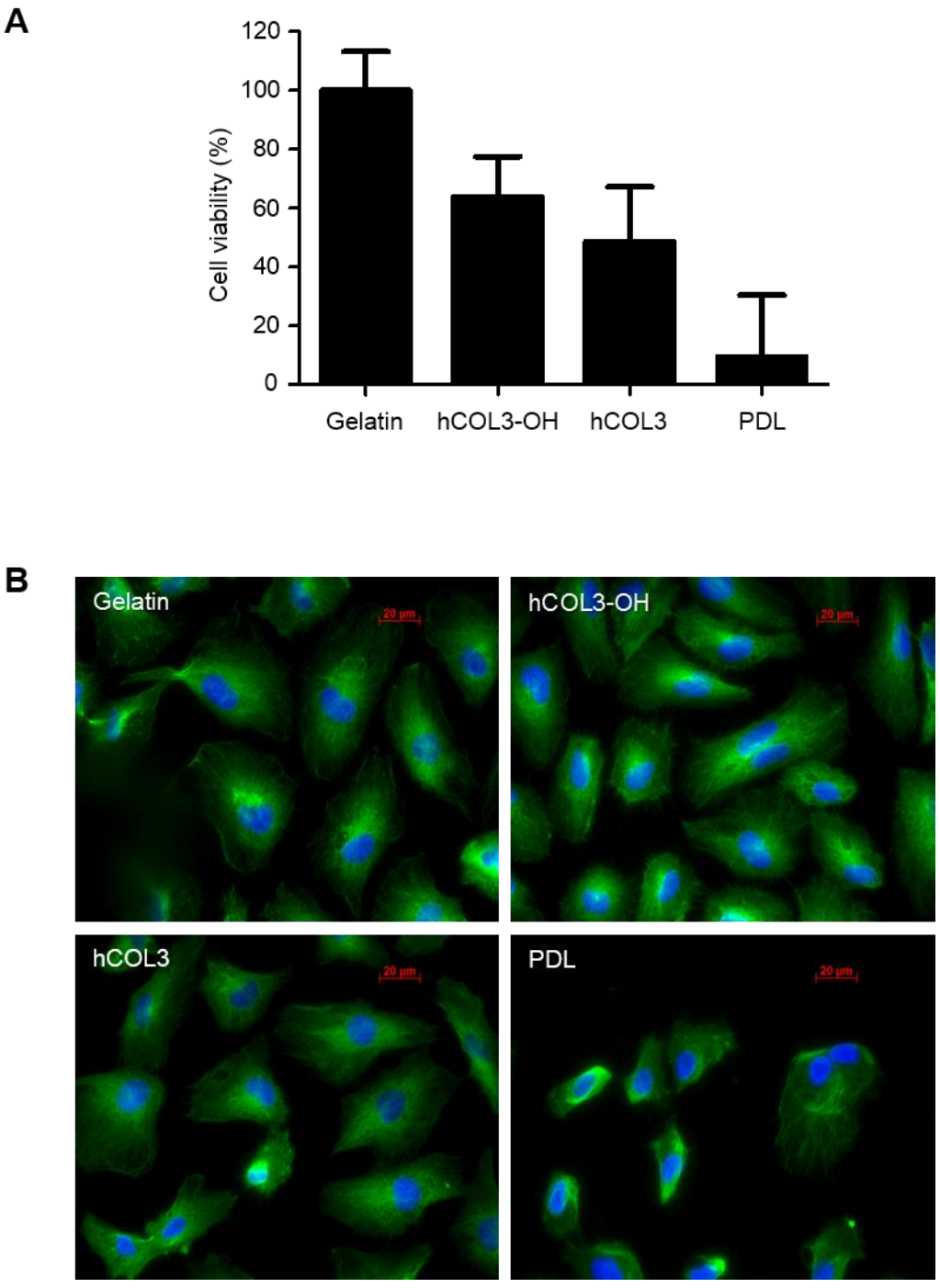

A
CD (mdeg)
Wavelength (nm)
B
% mdeg
Temperature (°C)
C
Trypsin
10 10 15 20 25 30 35 °C
250
130
100
70
55
35
25

FIGURE 7


A
120
100
80
60
40
20
0
Cell viability (%)
Gelatin
hCOL3-OH
hCOL3
PDL
B
Gelatin
hCOL3-OH
hCOL3
PDL
20 µm

REFERENCES (WITHOUT MANUSCRIPT AND PUBLICATION)

1. Heikkinen, J., et al., *Lysyl hydroxylase 3 is a multifunctional protein possessing collagen glucosyltransferase activity.* J Biol Chem, 2000. **275**(46): p. 36158-63.
2. Rautavuoma, K., et al., *Characterization of three fragments that constitute the monomers of the human lysyl hydroxylase isoenzymes 1-3. The 30-kDa N-terminal fragment is not required for lysyl hydroxylase activity.* J Biol Chem, 2002. **277**(25): p. 23084-91.
3. Vranka, J.A., L.Y. Sakai, and H.P. Bachinger, *Prolyl 3-hydroxylase 1, enzyme characterization and identification of a novel family of enzymes.* J Biol Chem, 2004. **279**(22): p. 23615-21.
4. Ishikawa, Y., et al., *Biochemical characterization of the prolyl 3-hydroxylase 1.cartilage-associated protein.cyclophilin B complex.* J Biol Chem, 2009. **284**(26): p. 17641-7.
5. Sethi, M.K., et al., *Identification of glycosyltransferase 8 family members as xylosyltransferases acting on O-glucosylated notch epidermal growth factor repeats.* J Biol Chem, 2010. **285**(3): p. 1582-6.
6. Sakaidani, Y., et al., *O-linked-N-acetylglucosamine modification of mammalian Notch receptors by an atypical O-GlcNAc transferase Eogt1.* Biochem Biophys Res Commun, 2012. **419**(1): p. 14-9.
7. Takeuchi, H., et al., *Rumi functions as both a protein O-glucosyltransferase and a protein O-xylosyltransferase.* Proc Natl Acad Sci U S A, 2011. **108**(40): p. 16600-5.
8. Sato, T., et al., *Molecular cloning and characterization of a novel human beta1,3-glucosyltransferase, which is localized at the endoplasmic reticulum and glucosylates O-linked fucosylglycan on thrombospondin type 1 repeat domain.* Glycobiology, 2006. **16**(12): p. 1194-206.
9. Schegg, B., et al., *Core glycosylation of collagen is initiated by two beta(1-O)galactosyltransferases.* Mol Cell Biol, 2009. **29**(4): p. 943-52.

GENERAL DISCUSSION

The association of heritable developmental disorders with defects in collagen modifying enzymes suggests that defects in glycosylating enzymes could be found in patients with connective tissue disorders. Should defects in genes for collagen glycosyltransferases prove to be the genetic cause of the patient's disease, these enzymes would be molecular targets for the development of possible therapies. Even though EDS, OI, and Marfan syndrome are genetic conditions and no cure exists, an improved understanding of the molecular mechanism provides information for personalized medical treatments. Such treatments aim at increasing bone mass and strength to prevent progressive deformities and fractures. Apart from medical interest, the production of recombinant collagen for biotechnological applications in tissue engineering, tissue remodeling after transplantation, and wound healing is in high demand. In order to fabricate tissues with characteristics closely resembling physiological tissues requires complete characterization of the physiological processes resulting in such tissues *in vivo*. Together, these points illustrate why identification of the ColGlcT is of considerable importance and interest as these findings should benefit many other fields in biomedicine and biotechnology.

Collagen glycans as diagnostic markers

Many collagen fragments derived from biosynthesis or degradation are cleared from the ECM by blood or urine drain. The presence of these remnants in body fluids renders them a valuable molecular target for diagnostic purposes. N-terminal propeptides freed from collagen type I or type III, PINP or PIIINP, reflect collagen biosynthesis, which can be assessed by measuring PINP or PIIINP in the blood by different assays [1]. Serum PINP is mostly affected by changes in bone metabolism with higher PINP levels in children than adults. In adults, serum PINP is a useful indicator for Paget's disease of bone [2], for early detection of metastasis to bone from breast, prostate, and lung cancer [3-5], and the degree and progression of osteoporosis [6]. Additional measures of bone formation are cross-linked N-telopeptidyl (NTx) and C-telopeptidyl (CTx) markers in urine and serum which are indicators of bone resorption and have slightly higher diagnostic sensitivity for bone metastasis than PINP. [7]. These cross-linked telopeptidyl markers can be liberated as peptide bound (NTx and CTx) or free forms. The free form exist as two different cross-linked structures, varying in the amount of Hyl forming the pyridinium cross-links. Pyridinolines (PYD or HP) are derived from three Hyl residues, whereas deoxypyridinolines (DPD or LP) originate from two Hyl and one Lys residue. Even though PYD and DPD are found in bone, cartilage, tendon, and dentin, urine PYD and DPD levels largely reflect the extent of bone resorption. In addition to providing information for diagnosing metabolic diseases, the PYD/DPD ratio is also used to address hydroxylation levels of collagen in hereditary diseases. In OI and EDS

under- or over- hydroxylation is reflected in the urine by the excretion of an abnormal PYD/DPD pattern, as observed in OI type IX and EDS type VIA. These markers are used mainly as a diagnostic tool to detect recessive OI types [8, 9].

Our screening of collagen glycosylation in cells from several untyped connective tissue disorder patients did not reveal defects in collagen glycosylating enzymes. Differences in the amount of glycosylation per α-chain are, however, associated with defective collagen biosynthesis as observed in OI and EDS patients [10]. Elevated glycosylation results from prolonged enzymatic activity during the slower assembly of defective collagen α-chains leading to more disaccharides but not to an increase in glycosylated sites [11]. Glycosylated PYD in urine has been associated with synovial tissue degradation in patients with joint diseases [12]. The distribution of glycosylated PYD among connective tissues is however, not clear and how glycosylated PYD changes during the course of a disease needs to be investigated.

Evaluating the sole contribution of PLOD3 for ColGlcT activity

Despite repeated identification of UGGT2 in ColGlcT active fractions, UGGT2 was not able to glucosylate Gal-based acceptor substrates such as denatured bovine type I collagen and *p*NP-Gal. In contrast, we could confirm the ability of human PLOD3 to glucosylate collagen type I and *p*NP-Gal to a low extent. PLOD3 is ubiquitously expressed with high expression early in development when basement membrane synthesis is important for the growing organism [13, 14]. The high expression level of PLOD3 in basement membrane associated tissue goes in hand with the highly glucosylated collagen types predominantly found in basement membranes [15, 16]. One possibility could be that PLOD3 is mainly glucosylating mesh-like collagen types as found in basement membranes or in the cuticle of *C.elegans*. Thus, a second enzyme would exist that preferentially glucosylates fibrillar collagen types. Furthermore the association of PLOD3 with basement membrane type IV collagen [16, 17] might indicate that PLOD3 could be the specific glucosyltransferase for non-fibrillar type collagen glycosylation. Considering the fact that the main collagen types in *C.elegans* are non-fibrillar basement membrane and cuticle types, *C.elegans* PLOD3 and GLT25D might be sufficient for collagen hydroxylation and glycosylation in *C.elegans*. Support for the theory of fibrillar and non-fibrillar type specific modification is the finding that in fibrillar collagen types the Lys residues in the helical domain are hydroxylated by PLOD1 and at the more distal telopeptidyl regions by PLOD2, making the LH activity of PLOD3 redundant in fibrillar type collagens.

The lack of PLOD3 in *G. gallus* is the strongest indicator for the existence of another enzyme with ColGlcT activity since the disaccharide structure is confirmed in *G. gallus* and the catalytic activity can be measured [18]. However, with the repeated identification of the gallus PLOD1 in ColGlcT active fractions, a possible contribution of PLOD1 to collagen glucosylation could be proposed.

Since *G. gallus* is the only animal that does not contain a PLOD3 or PLOD3-like protein, the question arises whether the avian PLOD3 isoform was evolutionarily lost or whether the chicken PLOD1 or PLOD2 retained a glycosyltransferase activity. Even though the gallus PLOD1 shares 76% identity with the human PLOD1 and only 59% with the human PLOD3, some N-terminal amino acid positions appear to be conserved between the gallus PLOD1, the human PLOD3, and the nematode PLOD enzymes (Table 1). Adjacent to the DDD motif are two Lys residues which are only shared between PLOD3 and the gallus PLOD1 but not the human PLOD1 or PLOD2. In the same region, the Ser at position 204 is also conserved in PLOD3, gallus PLOD1, and the nematode PLOD. A single Cys residue at position 144 in human PLOD3 was shown to be essential for ColGlcT activity, however, might not be directly involved in the active site [19]. This Cys residue is not conserved in gallus PLOD1. Whether all these amino acids are crucial for ColGlcT activity is not known. Nevertheless, human PLOD1 and PLOD2 expressed in insect cells do not possess glycosylating activities [20]. With gallus PLOD1 being more closely related to mammalian PLOD1 than PLOD3, a functional similarity shared only between gallus PLOD1 and human PLOD3 is highly doubtful.

Table 1: Protein sequence alignment of gallus PLOD isoforms with human PLOD isoforms and the nematode PLOD

Accession [a]	Protein	AA region 165 – 207 [b]	% identity to gPLOD1
gi|513199422|ref|XP_422695.4|	gallus PLOD2 X2	RIVQQWNLQDNDDDQLFYTKIYVDPLARERLNITLDHKCAIF	60.44
gi|513199419|ref|XP_004943367.	gallus PLOD2 X3	RIVQQWNLQDNDDDQLFYTKIYVDPLARERLNITLDHKCAIF	60.44
sp|O00469|PLOD2_HUMAN	human PLOD2	RIVQQWNLQDNDDDQLFYTKVYIDPLKREAINITLDHKCKIF	59.62
sp|Q02809|PLOD1_HUMAN	human PLOD1	KLVAEWEGQDSDSDQLFYTKIFLDPEKREQINITLDHRCRIF	76.48
gi|54111425|ref|NP_001005618.1	gallus PLOD1	KLVEEWKGKDDDSDQLFYTKIFLDPEKRENINISLDHRSRIF	100.00
sp|O60568|PLOD3_HUMAN	human PLOD3	QIVRQWKYKDDDDDQLFYTRLYLDPGLREKLSLNLDHKSRIF	58.93
sp|Q20679|PLOD_CAEEL	nematode PLOD	KILKLKSVEDKDDDQLYYTMIYLDEKLRKELNMDLDSMSKIF	43.65
		::: . :*.*.***:** :::* *: :.: ** . **	

[a] retrieved from NCBI upon pBlast of human PLOD3 against *G.gallus* RefSeq

[b] amino acid (AA) region homologue to human PLOD1 region 165 – 207

Sequence alignment with EMBL-EBI ClustalW2

To this extent, the mimiviral enzyme L230 which contains an LH domain and a GT domain supports the existence of such multifunctional proteins and evokes the interesting question of how these enzymes evolved. Between the three domains of life mimivirus is most closely related to eukaryotes [21]. Whether PLOD3 is a remnant of a bifunctional ancestor gene which has been separated into the two distinct LH (PLOD) and GT (GLT25D) enzymes or has evolved differently from the viral enzyme is unknown.

Glc-Gal-Hyl and ColGlcT as ligand and receptor in the ECM

Since the pioneering works of the seventies, little progress has been made investigating the function of collagen glycosylation. Several studies in the recent past assigned the function of collagen glycosylation mainly to collagen's own biosynthesis and regulation [22-24]. In my personal opinion, the potential for the glycan structure to serve as ligand for surrounding cells could be considered the most likely function for the collagen glycans. Glycosylation plays

important roles in cell-cell contacts, adhesion, migration, homing, virus binding, cancer biology and autoimmunity [25-27]. All these receptor-ligand functions arise from the fact that glycan structures decorate surfaces, inevitably causing them to be the first target recognized upon the initial contact of an interaction. Recently, Stawikowski and coworkers described a mechanism for how collagen might modulate integrin based adhesion of cells to the basement membrane. Glycosylated Hyl residues facilitate integrin-binding in a dose dependent manner [28]. The association of glycosylated Hyl residues with receptors recruiting cells to the basement membrane has been investigated as early as the late seventies. The identification of collagen modifying enzymes, in particular the ColGlcT, in platelets that lack collagen-like proteins in their membranes provoked the theory of ColGlcT mediated platelet-collagen interaction [29-31]. The platelet expressed membrane-bound ColGlcT has been thought to recognize Gal-Hyl structures on native triple helical collagen. Upon interaction, ColGlcT mediates platelet aggregation forming the collagen containing haemofibrotic plug. However, the ColGlcT mediated platelet interaction theory could not be confirmed but rather was dismissed since native triple helical collagen is not a substrate for ColGlcT [32, 33] and the ColGlcT activities have since been found to be located in the soluble platelet and plasma fractions [34-36]. Platelet activation and adhesion mediated by collagen is now thought to occur via GPVI receptor and integrin α2β1 [37]. Yet, the possible secretion of ColGlcT into blood plasma still remains puzzling. Differentially localized ColGlcT activities, namely ER-resident and secreted [36, 38] supports the possibility of more than one gene coding for ColGlcT activity.

Prospective research targets for collagen glycosylation

Since siRNA degradation of specific enzymatic activities are never 100%, the application of a new biochemical tool, the CRISPR/Cas-system to specifically disrupt a gene of interest might be effective in eliminating PLOD3 function in human fibroblasts. By eliminating PLOD3 function, a possible contribution of another enzyme to collagen glucosylation could be investigated. This could answer not only whether PLOD3 is sufficient for collagen glucosylation, but could address what impact the missing Glc has on proper collagen folding or secretion. The CRISPR/Cas-system could also be used to investigate the general contribution of collagen glycosylation towards collagen synthesis, folding, and secretion.

Another biochemical approach to address the function of the collagen glycan could include production of recombinant hydroxylated collagen in glycosylated and un-glycosylated forms. The Mimivirus L230 enzyme has been found to add a Glc residue to the polypeptide Hyl producing a Glc-Hyl glycan structure rather than the animal Gal-Hyl or Glc-Gal-Hyl structures [39]. Since we have successfully introduced the mimiviral hydroxylases L593 and L230 into a bacterial expression system to produce hydroxylated collagen [40], the production of glycosylated collagen

would be the next step towards the production of physiological collagen. Site-directed mutagenesis studies in the GLT25D1 galactosyltransferase revealed two essential DXD motifs important for the catalytic function [41]. The region with these two motifs could serve as a guiding principle for site directed mutation of the glucosyltransferase domain of L230 to shift from a glucosyltransferase activity to a galactosyltransferase activity. As yet, the conversion of a glycosyltransferase's donor specificity from a UDP-hexose to another UDP-hexose has not been reported, hence the exact sequence information determining the donor specificity is unknown and its description would be a pioneering work. Since the conversion of the donor specificity of a glycosyltransferase is an ambitious task to achieve, another possibility to produce glycosylated collagen in *E.coli* could be the coexpression of the *C.elegans* ColGalT with collagen and the mimiviral hydroxylases. The nematode enzyme is not glycosylated itself and is active in bacteria unlike the mammalian ColGalT (unpublished data).

While the biochemical approaches to identify the ColGlcT did not provide clear answers, the bioinformatics approach revealed the candidate enzyme GTDC1 that should be considered a strong candidate for a collagen glucosyltransferase. GTDC1 is widespread among the animal kingdom, belongs to the retaining GT-families for α-glycosyltransferases, contains an ER-retention signal sequence, and has no known function assigned to it. Additionally of interest, proteomic analysis of collagen mediated platelet aggregation revealed the existence of GTDC1 in platelets among other collagen modifying enzymes, including PLOD1 and PLOD3 [42]. Conclusively, GTDC1 should be considered to examine for collagen glucosyltransferase activity.

References

1. Koivula, M.K., L. Risteli, and J. Risteli, *Measurement of aminoterminal propeptide of type I procollagen (PINP) in serum.* Clin Biochem, 2012. **45**(12): p. 920-7.
2. Alvarez, L., et al., *Components of biological variation of biochemical markers of bone turnover in Paget's bone disease.* Bone, 2000. **26**(6): p. 571-6.
3. Haider, M.T., et al., *Modifying the osteoblastic niche with zoledronic acid in vivo-potential implications for breast cancer bone metastasis.* Bone, 2014. **66**: p. 240-50.
4. Valencia, K., et al., *miR-326 associates with biochemical markers of bone turnover in lung cancer bone metastasis.* Bone, 2013. **52**(1): p. 532-9.
5. Koopmans, N., et al., *Serum bone turnover markers (PINP and ICTP) for the early detection of bone metastases in patients with prostate cancer: a longitudinal approach.* J Urol, 2007. **178**(3 Pt 1): p. 849-53; discussion 853; quiz 1129.
6. Ebeling, P.R., J.M. Peterson, and B.L. Riggs, *Utility of type I procollagen propeptide assays for assessing abnormalities in metabolic bone diseases.* J Bone Miner Res, 1992. **7**(11): p. 1243-50.
7. Joerger, M. and J. Huober, *Diagnostic and prognostic use of bone turnover markers.* Recent Results Cancer Res, 2012. **192**: p. 197-223.

8. Kraenzlin, M.E., et al., *Automated HPLC assay for urinary collagen cross-links: effect of age, menopause, and metabolic bone diseases.* Clin Chem, 2008. **54**(9): p. 1546-53.
9. Rohrbach, M. and C. Giunta, *Recessive osteogenesis imperfecta: clinical, radiological, and molecular findings.* Am J Med Genet C Semin Med Genet, 2012. **160C**(3): p. 175-89.
10. Cabral, W.A., et al., *Abnormal type I collagen post-translational modification and crosslinking in a cyclophilin B KO mouse model of recessive osteogenesis imperfecta.* PLoS Genet, 2014. **10**(6): p. e1004465.
11. Taga, Y., et al., *Site-specific quantitative analysis of overglycosylation of collagen in osteogenesis imperfecta using hydrazide chemistry and SILAC.* J Proteome Res, 2013. **12**(5): p. 2225-32.
12. Gineyts, E., P. Garnero, and P.D. Delmas, *Urinary excretion of glucosyl-galactosyl pyridinoline: a specific biochemical marker of synovium degradation.* Rheumatology (Oxford), 2001. **40**(3): p. 315-23.
13. Wang, C., M. Valtavaara, and R. Myllylä, *Lack of collagen type specificity for lysyl hydroxylase isoforms.* DNA Cell Biol, 2000. **19**(2): p. 71-7.
14. Salo, A.M., et al., *The lysyl hydroxylase isoforms are widely expressed during mouse embryogenesis, but obtain tissue- and cell-specific patterns in the adult.* Matrix Biol, 2006. **25**(8): p. 475-83.
15. Myllylä, R., et al., *Expanding the lysyl hydroxylase toolbox: new insights into the localization and activities of lysyl hydroxylase 3 (LH3).* J Cell Physiol, 2007. **212**(2): p. 323-9.
16. Ruotsalainen, H., et al., *Glycosylation catalyzed by lysyl hydroxylase 3 is essential for basement membranes.* J Cell Sci, 2006. **119**(Pt 4): p. 625-35.
17. Rautavuoma, K., et al., *Premature aggregation of type IV collagen and early lethality in lysyl hydroxylase 3 null mice.* Proc Natl Acad Sci U S A, 2004. **101**(39): p. 14120-5.
18. Myllyla, R., et al., *Isolation of collagen glucosyltransferase as a homogeneous protein from chick embryos.* Biochim Biophys Acta, 1977. **480**(1): p. 113-21.
19. Wang, C., et al., *Identification of amino acids important for the catalytic activity of the collagen glucosyltransferase associated with the multifunctional lysyl hydroxylase 3 (LH3).* J Biol Chem, 2002. **277**(21): p. 18568-73.
20. Heikkinen, J., et al., *Lysyl hydroxylase 3 is a multifunctional protein possessing collagen glucosyltransferase activity.* J Biol Chem, 2000. **275**(46): p. 36158-63.
21. Suzan-Monti, M., B. La Scola, and D. Raoult, *Genomic and evolutionary aspects of Mimivirus.* Virus Res, 2006. **117**(1): p. 145-55.
22. Terajima, M., et al., *Glycosylation and cross-linking in bone type I collagen.* J Biol Chem, 2014. **289**(33): p. 22636-47.
23. Parisuthiman, D., et al., *Biglycan modulates osteoblast differentiation and matrix mineralization.* J Bone Miner Res, 2005. **20**(10): p. 1878-86.
24. Pokidysheva, E., et al., *Posttranslational modifications in type I collagen from different tissues extracted from wild type and prolyl 3-hydroxylase 1 null mice.* J Biol Chem, 2013. **288**(34): p. 24742-52.
25. Ohtsubo, K. and J.D. Marth, *Glycosylation in cellular mechanisms of health and disease.* Cell, 2006. **126**(5): p. 855-67.
26. Moremen, K.W., M. Tiemeyer, and A.V. Nairn, *Vertebrate protein glycosylation: diversity, synthesis and function.* Nat Rev Mol Cell Biol, 2012. **13**(7): p. 448-62.
27. Maverakis, E., et al., *Glycans in the immune system and The Altered Glycan Theory of Autoimmunity: a critical review.* J Autoimmun, 2015. **57**: p. 1-13.

28. Stawikowski, M.J., et al., *Glycosylation modulates melanoma cell alpha2beta1 and alpha3beta1 integrin interactions with type IV collagen.* J Biol Chem, 2014. **289**(31): p. 21591-604.
29. Barber, A.J. and G.A. Jamieson, *Platelet collagen adhesion characterization of collagen glucosyltransferase of plasma membranes of human blood platelets.* Biochim Biophys Acta, 1971. **252**(3): p. 533-45.
30. Jamieson, G.A., C.L. Urban, and A.J. Barber, *Enzymatic basis for platelet: collagen adhesion as the primary step in haemostasis.* Nat New Biol, 1971. **234**(44): p. 5-7.
31. Katzman, R.L., A.H. Kang, and E.H. Beachey, *Collagen-induced platelet aggregation: involement of an active glycopeptide fragment (alpha1-CB5).* Science, 1973. **181**(4100): p. 670-2.
32. Myllylä, R., L. Risteli, and K.I. Kivirikko, *Glucosylation of galactosylhydroxylysyl residues in collagen in vitro by collagen glucosyltransferase. Inhibition by triple-helical conformation of the substrate.* Eur J Biochem, 1975. **58**(2): p. 517-21.
33. Anttinen, H., et al., *Intracellular enzymes of collagen biosynthesis in human platelets.* Blood, 1977. **50**(1): p. 29-37.
34. Menashi, S., R. Harwood, and M.E. Grant, *Native collagen is not a substrate for the collagen glucosyltransferase of platelets.* Nature, 1976. **264**(5587): p. 670-2.
35. Menashi, S. and M.E. Grant, *Studies on the collagen glucosyltransferase activity present in platelets and plasma.* Biochem J, 1979. **178**(3): p. 777-84.
36. Leunis, J.C., et al., *The distribution of collagen:glucosyltransferase in human blood cells and plasma.* Biochim Biophys Acta, 1980. **611**(1): p. 79-86.
37. Ruggeri, Z.M., *Platelets in atherothrombosis.* Nat Med, 2002. **8**(11): p. 1227-34.
38. Henkel, W. and E. Buddecke, *Purification and properties of UDP-glucose galactosylhydroxylysine collagen glucosyltransferase (EC 2.4.1.?) from bovine arterial tissue.* Hoppe Seylers Z Physiol Chem, 1975. **356**(6): p. 921-8.
39. Luther, K.B., et al., *Mimivirus collagen is modified by bifunctional lysyl hydroxylase and glycosyltransferase enzyme.* J Biol Chem, 2011. **286**(51): p. 43701-9.
40. Rutschmann, C., et al., *Recombinant expression of hydroxylated human collagen in Escherichia coli.* Appl Microbiol Biotechnol, 2014. **98**(10): p. 4445-55.
41. Perrin-Tricaud, C., C. Rutschmann, and T. Hennet, *Identification of domains and amino acids essential to the collagen galactosyltransferase activity of GLT25D1.* PLoS One, 2011. **6**(12): p. e29390.
42. Milioli, M., et al., *Quantitative proteomics analysis of platelet-derived microparticles reveals distinct protein signatures when stimulated by different physiological agonists.* J Proteomics, 2015. **121**: p. 56-66.

ACKNOWLEDGEMENTS

At the very beginning, I would like to thank all the people who contributed in various manners to the completion of my PhD thesis. I thank my supervisor Prof. Dr. Thierry Hennet for giving me the opportunity to accomplish my PhD studies in your research group. And in particular I am very thankful that you always had an open door, were constantly available to address scientific questions, and for your great deal of patience.

I would like to express thanks to the members of the thesis committee PD Dr. Lubor Borsig, Prof. Dr. Arnold von Eckardstein, and Prof. Dr. Matthias Baumgartner for their time and effort to analyze and review the present research work and for helpful suggestions during the committee meetings.

During my thesis I came sometimes to a point where I needed external assistance. I would like to thank Dr. Peter Gehrig and Dr. Simon Barkow from the Functional Genomic Center Zürich for your huge help to obtain and analyze masspectrometric data.

Then, I would like to acknowledge all the excellent researchers I was working with in the lab: Christoph Rutschmann for teaching me a lot of labtechniques, for our awesome collaboration, and for all the mental support during the quick ristretto breaks. Anna Rommel for correcting my funny cowboy writing style, for your immense support at the finish line, and all your delicious cupcakes in the coffee breaks, omnom. Dr. Kelvin Luther, for having always an idea how to continue with my work and in the rare cases you did not, your sense of humor brought me back on track with a big smile☺, GoWingsGo. Dr. Andreas Hülsmeier for all the fruitful scientific discussions and for sharing your deep and profound biochemical knowledge with me, which also fed my fascination for masspectrometry technology. Dr. Nikunj Shah, the alltime bestest inder and 'dear friend' for being an awesome bench neighbor in the bestest cloning lab. Dr. Adrienne Weiss for the sweet side of life, no not glycans, cakes and happiness. Dr. Michael Welti, Yen-lin Eddie Huang, and Sacha Schneeberger for your Asian style I very much appreciated as in Jiu Jitsu lessons, Chinese lessons, and all kinds of peculiar foods you brought, mjamm. Giovanna Roth, for the early morning coffee chats and the awesome pasta dinners again and yet again at your place. Furthermore I would like to thank former and current lab members Dr. Andrea Fuhrer, Stephan Baumann, Marek Whitehead, Nina Hochhold, Luca Plan, and the Borsigs Dr. Irina Häuselmann, Marko Roblek, Jesus Glaus Garzón (for the many jesusito hugs ☺), Darya Protsyuk, and Esther Quinzano.

Since I owe the accomplishment of this thesis not only to scientists, I wish to express my deep gratitude to my parents, Heidi and Andrin, my sister Dr. Katrin, my brother Urs, Stephi, my friends, and my girlfriend Anna for patient and tremendous mental support and encouragement in so many different ways. Thank you all!!!